Using a dynamic exposure model to improve understanding of exposure to urban air pollution

Smith, James David

Awarding institution:
King's College London

Using a dynamic exposure model to improve understanding of exposure to urban air pollution

James D Smith, MSc, BA, PgCert, PgDip

Environmental Research Group, School of Analytical
& Environmental Sciences, Faculty of Life Sciences
& Medicine, King's College London

Thesis submitted to King's College London in fulfillment of the requirements for the degree of Doctor of Philosophy

December 2018

Abstract

Epidemiological studies of the negative effects on health of poor air quality are typically based on subjects' residential address. These 'static' methods may be assigning exposure to subjects/populations incorrectly. Possible sources of error include the coarse spatial and temporal scale of the pollutant data, failing to account for lack of movement of the subjects, and not adequately modelling the effects of microenvironments. This PhD takes a large Transport for London (TfL) survey (the 'LTDS') of Londoners daily activities and uses geographical information science techniques to create a detailed model (the 'LTDS-X') of Londoners typical movements including time of day, location and microenvironment. This model is then combined with the Kings version of the Community Multiscale Air Quality model (CMAQ-Urban), which is a multi-pollutant and multi-source high resolution spatial and temporal model of UK air quality. By combining the LTDS-X with CMAQ-Urban and then undertaking further micro-environmental modelling on top of this (in-car, in-train, indoors, the London Underground) detailed exposure estimates to $PM_{2.5}$ and NO_2 for the population of London are calculated and then compared to the 'static' exposure method. Results show that exposure indoors, and whether or not subjects use the London Underground (for $PM_{2.5}$ exposure), were important determinants of Londoners daily exposure. The $PM_{2.5}$ exposure modelling for when subjects were on the London Underground was therefore investigated further with a measurement campaign across the network, resulting in a spatial routing model of the network ('TubeAir'). As a stand-alone model this will be useful for future exposure studies in London, and its use was demonstrated on a sample journey. This research concludes by exploring the difficulty of evaluating hybrid exposure models in terms of the representativeness of any exposure calculated, by comparing measured concentrations on a repeated number of cycling journeys with modelled exposures on the same journey.

n.b. The LaTeX files for this thesis are available at : `https://github.com/JimShady/PhD`

Dedication

For Aimee and Lola.

Acknowledgements

I would like to express gratitude to my supervisors, Dr Benjamin Barratt and Dr Heather Walton. Their guidance, ideas and support have been crucial. I would also like to thank Professor Frank Kelly. who gave me the opportunity to start this part-time PhD program in the first place. In may 'day job', current and previous work colleagues in our group have helped me at various times, mainly Dr Sean Beevers, Dr Christina Mitsakou, Dr Nutthida Kitwiroon, Dr Andrew Beddows, Dr Gregor Stewart, Dr David Dajnak, Dr David Green, Dr Ian Mudway and Dr Gary Fuller. University aside, special thanks go to my family, friends and in particular my wife Aimee for putting up with my elevated stress levels for so long. Also to Frankie Manning, for giving me a way to express myself in a totally different way.

Finally, a little thank you to the various online communities that I have interacted with and sought advice from over the years, who were able to offer advice and suggestions when I ran into practical problems that nobody else could solve. I try to give back when I can.

Declaration

I, James David Smith, declare that all the work submitted in this thesis is my own and that all references are cited accordingly.

Signed: _____

Date: _____

Contents

List of Figures

List of Tables

1. Introduction

Air pollution, and it's related health impacts, is an issue that urban areas have been struggling with for many years. In Roman times the philosopher Seneca commented (after modern translation) that his disposition improved after leaving the heavy air of Rome (Seneca and Campbell (1969)) and after large cities in the UK started to use coal as a fuel in the 13th century, the wife of Henry VIII complained of coal smoke in the air when she visited Nottingham Castle (Brimblecombe (1999)). The links between air pollution and negative effects on human health are still with us today, although the sources have changed and the visibility of the pollution has in many cases become less apparent. The Lancet's Global Burden of Disease 2013 study ranked outdoor air pollution as the 7th highest contributing risk factor to deaths globally in 2010 (Figure 1.1) (Lim et al (2012)).

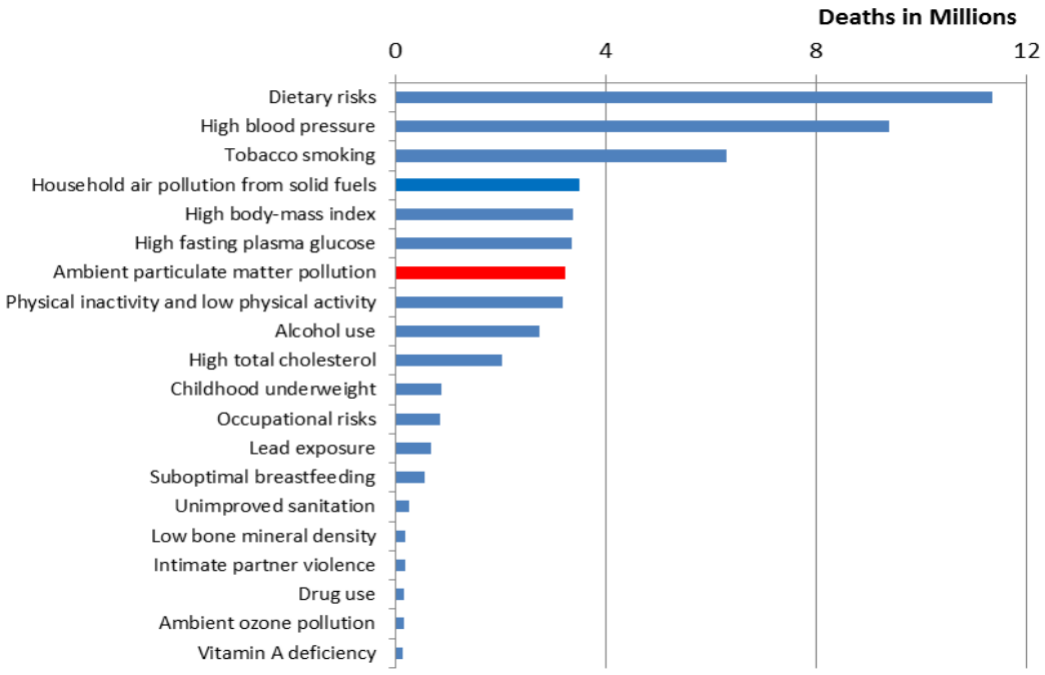


Figure 1.1: 20 leading risk factors contributing to deaths globally in 2010

However the methods by which exposure to air pollution is estimated are under constant revision. Due to limitations in these methods, epidemiologists may therefore be making misleading or incorrect conclusions.

This PhD begins by giving a general introduction to the subject of air pollution, it's health effects, and how health studies of air pollution have estimated population exposure in the past. It then conducts a more detailed review of dynamic exposure-health studies and the current 'state of play'. This sets the scene for novel research into the personal exposure of the population of London, and the understanding of the results thereof.

2. Background

2.1 Air pollution - sources and behaviour

2.1.1 What is air pollution?

Air pollution is defined by Colls (1997) as "material emitted into the air from stationary or mobile sources, moving subsequently through an aerial path and perhaps being involved in chemical or physical transformation before eventually being returned to the surface". This research focuses on the places in which humans, particularly in urban environments, are exposed to this pollution.

2.1.2 Particle types and sizes

Air pollution is a summary term for many sub-categories of pollutants. Pollutants can be solid particles, liquid particles, or gaseous material. They can also be classified as either a primary or secondary pollutant. For example fumes emitted from the stack of a power station are classified as primary pollutants, whereas ground level ozone formed by chemical reactions between primary pollutants catalysed by sunlight are classed as secondary pollutants. The UK Department for Environment, Food and Rural Affairs (DEFRA) is 3 hours behind summarises the main constituents of air pollution and their typical sources as follows (DEFRA (2011a)):

- Particulate matter
  - Combustion (traffic or stationary sources), sea-spray, construction, quarrying. 'Fine' refers to particulate matter smaller than 1 micron in diameter (PM_1), ultrafine smaller than 2.5 microns ($PM_{2.5}$, includes the 1 micron particles), and coarse smaller than 10 microns (PM_{10}, includes the 1 and 2.5 micron particles).
- Oxides of nitrogen (NO_x)
  - Combustion. Road transport, electrical supply industry, other industry.

- Ozone (O_3)
  - A secondary pollutant, not emitted directly from human-made sources, but formed as a result of reactions between other pollutants (Oxides of nitrogen, volatile organic compounds) in sunlight.
- Sulphur dioxide ($SO2_2$)
  - Combustion of fuels such as coal and heavy oils by power stations.
- Polycyclic aromatic hydrocarbons (PAHs)
  - Many different sources. DEFRA uses Benzo[a]pyrene as a marker. Main sources are coal and wood burning, fires, industrial processes. Traffic combustion (diesel in-particular) is a major contributor.
- Benzene
  - Domestic and industrial combustion and road transport.
- 1,3-butadiene
  - Combustion of petrol i.e. motor vehicles that use petrol as a fuel source
- Carbon monoxide (CO)
  - Occurs from incomplete combustion of fuels that contain carbon. Road transport, residential combustion and industrial combustion are the main sources.
- Lead (Pb)
  - Combustion of coal and nonferrous metals
- Ammonia
  - Mainly from agriculture such as manure, fertilisers and slurry.

When discussing the amount of pollutants in the air, either volumetric or gravimetric units are used. Volumetric units quantify the ratio of volume of pollutants to clean dry air (itself a mixture of nitrogen, oxygen, argon etc.), whereas gravimetric units quantify the mass of the material per volume of air. Most laws and guidelines, such as the European Union Air Quality Standards (see Section 2.2), use gravimetric measurements. Table 2.1 summarises the abbreviations for volumetric and gravimetric units which are used throughout this research.

Table 2.1: Abbreviations for volumetric and gravimetric pollutant units

Volumetric	
Description	**Notation**
Parts per million of pollutant per parts of air volume	(10^{-6} ppm)
Parts per billion of air pollutant per parts of air volume	(10^{-9} ppb)
Parts per trillion of air pollutant per parts of air volume	(10^{-12} ppt)
Gravimetric	
Description	**Notation**
milligrams of pollutant per cubic metre	(mg/m^3)
micrograms of pollutant per cubic metre	($\mu g\ m^{-3}$)
nanograms of pollutant per cubic metre	(ng/m^3)

Of the ten pollutants listed above, over half list transport combustion as being a source. When combined with proximity to humans in urban environments it is easy to see why air quality in towns and cities, and in particular the pollutants caused by vehicles in these environments, receives such great interest in the field of air quality research and environmental science.

2.1.3 Urban environments

In many cities around the world hundreds of thousands of people now live within metres of major pollution sources such as car-filled roads, power stations or industrial plants. According to the World Health Organisation (WHO), as of 2010, more than 50% of the worlds population live in urban areas. This is up from 40% in 1990. The prediction is that by 2050 this number will rise to 70% (Global Health Observatory (2012)). As the numbers of people living in cities has grown, so has the infrastructure required to support them; much of which causes air pollution. High numbers of people are now being exposed to air pollution above WHO guidelines during their normal day to day activities. Given this close link between population density and pollution, it is important to understand the complexity of air pollution in urban environments.

In 2012 WHO published a review which summarised data on particulate matter levels in major cities across the world. The data were split into two categories, particulate matter of a diameter of less than 2.5 micrometres (referred to as $PM_{2.5}$) and particulate matter of a diameter of less than 10 micrometres (referred to as PM_{10}) (World Health Organization (2012)). Although the health effects of poor air quality will be discussed in Section 2.2, to provide immediate context, we can refer to WHO Factsheet no. 313 which gives the following numbers as limits for 'acceptable and achievable objectives to minimize health effects' (World Health Organization (2011)). Nitrogen dioxide (NO_2) values are also included for

future reference.

Table 2.2: Table of WHO PM and NO_2 objectives

	Annual mean (μg m^{-3})	24 hour mean (μg m^{-3})
$PM_{2.5}$	10	25
$PM_{2.5}$	20	50
NO_2	40	200

Figure 2.1 shows the annual average PM_{10} levels for each major world city for 2003-2010, weighted by population, from the same WHO review.

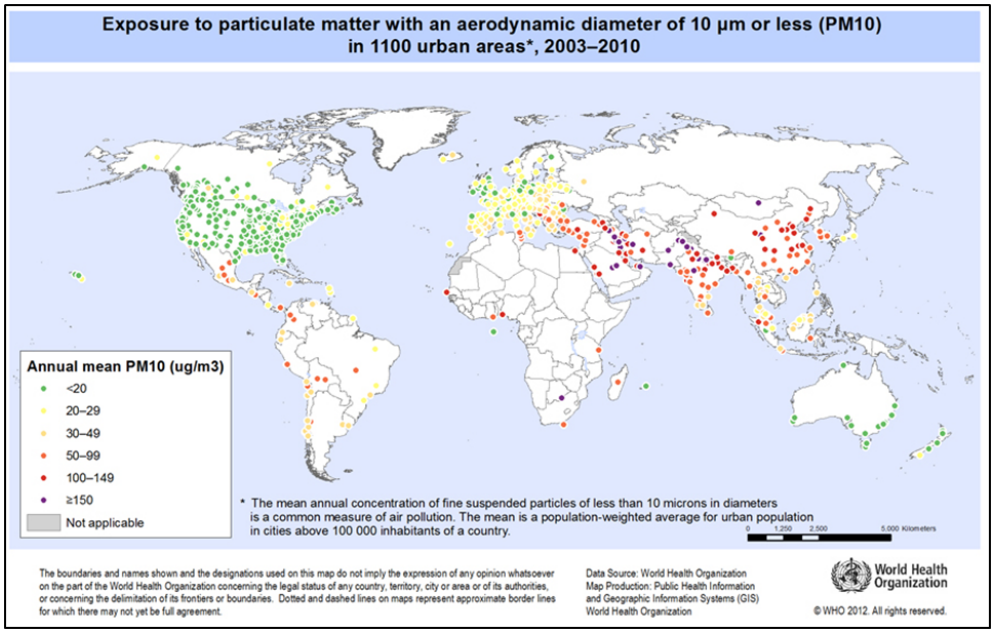


Figure 2.1: A map of PM_{10} in major world cities

As can be seen, there are many urban areas where the WHO annual mean for PM_{10} is exceeded. By way of example we can consider Beijing which in 2013 had a population of 15.59m (The United Nations Statistics Division (2013)) and was China's second largest city, is subject to seasonal dust storms, hot humid summers and cold dry winters - and suffers from severe air pollution problems. During the 1990s attempts were made at controlling air pollution in the city by introducing the use of low-sulphur coal, using natural gas as an alternative to coal, phasing out leaded petrol, and moving factories and heavy industry outside of the city. However in the early 00s, due to increasing vehicle numbers and the rapid growth of the industrial sector, particle levels continued to remain higher than national standards (SUN et al. (2004)). The nearby area of Hebei Province (which rings Beijing and is heavily industrial) made the efforts of the government to control air quality in Beijing at this time even more difficult as the Hebei region has lower fuel quality standards, and

can cause some primary emissions to drift into Beijing (Tuo et al. (2013)) due to prevailing winds (Meteorology is discussed in Section 2.1.4).

During the early 21^{st} century the use of air pollution monitoring equipment became more commonplace and better data was collected to enable scientists to understand the issues that Beijing (and similar urban environments) faced. SUN et al. (2004) found that during the years 1999 and 2000, ambient $PM_{2.5}$ concentrations were in the range 37 – 357 μg m^{-3} , and estimated a yearly average of 89.7 μg m^{-3}. The research concluded that coal burning and traffic exhausts, along with dust from long range sources, were the major pollution sources in the urban environment of Beijing (SUN et al. (2004)).

The study of pollution in urban environments is essential as these areas are where humans are most readily exposed and they are where the sources of emissions are most frequent. Although PM has been discussed here, similar issues apply to other traffic linked pollutants such as NO_x and PAHs. There are also other factors, other than the type of pollutant, that complicate the understanding of pollution in urban environments, such as weather and geography. Within these urban environments there are hyper-local conditions that can raise and lower levels. These notions are explored in sections 2.1.4, 2.1.5, 2.1.6 and 2.1.7.

2.1.4 Meteorology

Local weather conditions have a strong influence on air quality. Air pollution can be removed from the air in the process of cloud formation, and then deposited on the ground when the clouds turn to rain at a future time and/or place. Falling rain can also remove pollutants from the air by collecting the pollution and 'cleaning' the air as it falls. Both of these processes are grouped into the term 'wet deposition'. A 6 μg m^{-3} difference in PM_{10} was observed in Edinburgh between days with no rainfall compared to those with more than 20mm of rain (DEFRA (2007)). In December 2013 it was even reported that China was considering using 'cloud seeding', i.e. the process of engineering the weather to rain, as a method to lower air pollution in the most polluted regions of China (Slezak (2013)).

Wind can adversely affect air quality by trapping or recirculating pollutants (discussed in Section 2.1.5), but can also disperse the pollutants or move them to other areas/regions. This notion was first proposed in the 1960s when studying the acidification of lakes in Scandinavia Summers and Whelpdale (1976), where it was theorised that the high acid levels were due to air quality elsewhere in Canada. In the UK NO_x ambient concentrations were found to have halved at a monitoring station in Hillingdon, United Kingdom (UK) when wind speed rose from 5 to 10 m/s-1, while $PM_{2.5}$ also decreased, however the coarse PM ($PM_{2.5}$ to PM_{10}) increased due to re-suspension of particles that had previously settled (DEFRA

(2007)). Depending on the lifetime and properties of a pollutant, it can be transported on scales ranging from the street level to the global scale (Monks et al. (2009)). Stohl (2003) found gases were being transported from North America to Europe. More locally, an odour event in the South-East of England on 18 April 2008 (The Guardian (2008)) was found to have originated from agricultural emissions in northern Germany (Smethurst et al. (2012)). The Geneva Convention on Long-range Trans-boundary Air Pollution was established in 1979 to look at ways to deal with this movement of air pollution between borders in terms of national air quality guidelines and limits, and came into force in 1983 (United Nations Economic Commission for Europe (1983)).

Direct sunlight (ultra-violet radiation) and higher temperatures, on hot summer days, can initiate reactions with nitrogen dioxide which can lead to the formation of ozone. The ozone and ozone forming chemicals remain in the atmosphere and can be transported over regional and national borders. This layer can then settle over cities such as London and lead to what is often referred to as summertime 'smog'. The South-East of England often has high concentrations during spring and summer as, amongst other sources, it is close to European pollution sources (King's College London (2013)). Although vehicle emissions of nitrogen oxide can have the effect of reacting with the ozone to lower the level of ozone in areas where vehicle emissions are particularly high. (Environmental Protection Agency (2012)).

2.1.5 Urban topography

In the urban environment, streets are often bordered by tall buildings which can influence pollution levels for people at ground level. Topography of this nature is often referred to as a 'street canyon'. In extreme circumstances such as on the streets of Hong Kong, skyscrapers are littered throughout the city, but on smaller scales such as Oxford Street (London, UK) similar issues occur but with modest building heights. Being bordered by tall buildings creates a sheltering effect from the wind, stopping particles being moved elsewhere. The typical street-canyon effect occurs when there is a steady flow over the top of tall buildings (see figure 2.2), where the mean flow is perpendicular to the direction of the street (Britter and Hanna (2003)). With roof-level wind-speeds of 1.5-2 m/s, the air is recirculated within this 'box' and air quality deteriorates as sources (usually traffic) emit more fumes.

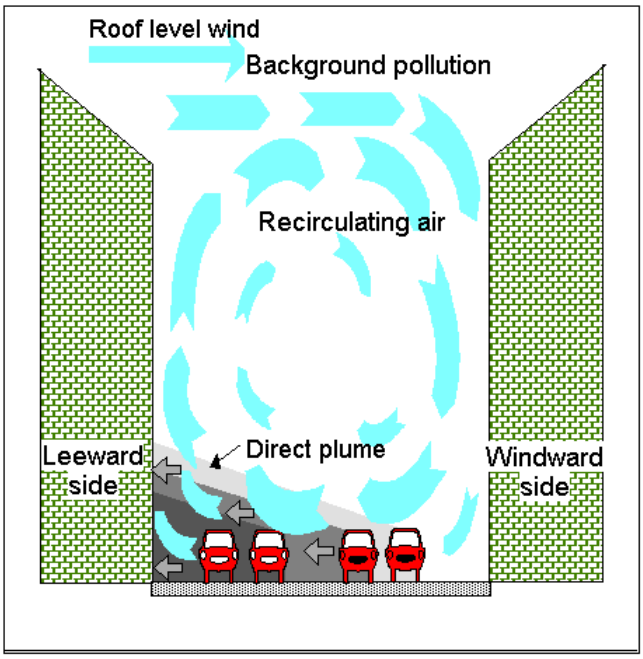


Figure 2.2: Pollutant dispersion in a regular street canyon

Mexico City is an example of a city that is subject to this effect on a large scale. Situated in central Mexico, North America (19.4328 N, 99.1333 W), Mexico City is one of the most populated cities in the world and has an estimated population of about 21m (2011) living within the Mexico City Metropolitan Area (MCMA) of 1,485 km^2 (The United Nations Statistics Division (2013)). Mexico City is located in the crater of a large extinct volcano, which means that the entire city suffers from the aforementioned canyon effect. Almost like it is surrounded by skyscrapers. This is exacerbated by a fleet of older vehicles with poor engines (the effects of which are discussed more in Section 2.1.7), low levels of oxygen (due to the high altitude of the city), and wind patterns that concentrate pollutants in the western and southern parts of the city (Garza (1996)) where the population is most dense.

To summarise, air pollution in the urban environment can be effected by the meteorology and topography of the area. On the smaller scale than 'area' there are also micro-environments that people spend time in, for example inside a bus, which can exhibit very different characteristics than the rest of the city or even street.

2.1.6 Micro-environments

Micro-environments are defined as the immediate small-scale environment of an organism, especially as a distinct part of a larger environment. Examples in the context of this the-

sis include the air quality inside a vehicle, a house or underground train carriage, in the context of the environment outside. Understanding how pollutant levels change in these micro-environments is key, as much of our time is spent within these environments, thus our exposure level to air pollutants whilst in them could have important impacts on our health.

2.1.6.1 Indoors

The most common micro-environment is within buildings, the air we are exposed to when we are at home, in the office or at school etc. WHO calculated in 2005 that people spend 89% of their time indoors (World Health Organization (2005)). In these environments, people are exposed to pollutants generated outdoors that penetrate to the indoor environment, as well as to pollutants produced indoors. The EXPOLIS study Kousa et al. (2002) examined how much time people spend in these environments by asking 1427 people from accross Europe (Athens (Greece), Basle (Switzerland), Grenoble (France), Helsinki (Finland), Milan (Italy). Oxford (Great Britain) and Prague (Czech)) to complete activity diaries. They found that the amount of time people spend indoors varied by whether the people were employed or not, in what type of job, whether they lived alone and/or whether they had children. Gender and season of year were also found to be factors. Study participants were found to spend on average 13.95 hours indoors at home (13.48 min to 15.76 max) and 6.71 hours indoors at work (5.09 to 7.09) (Schweizer et al. (2007)). So understanding indoor air quality, which would include filtration of outdoor air into the building as well as pollutants whose sources are inside, is important in understanding personal exposure. However research on indoor pollution has not had the same focus as outdoor pollution for a number of reasons. Firstly the perceived need to deal with coal and traffic emissions, the ease of monitoring outdoor air quality using fixed monitoring sites (compared to monitoring every home). Secondly epidemiologists have traditionally only linked outdoor ambient pollutant concentrations to health health issues, furthermore, legislating the air that people can breathe in their own homes can be seen as intrusive to people's private lives, finally the funding and policy initiatives around air pollution research has mainly come from developed countries, which who do not have such an issue with indoor pollution as low and middle income countries (World Health Organization (2010)) (due to low and middle income countries using solid fuel for cooking and heating inside their homes more).

The pollutants that are emitted indoors, and which the US Environmental Protection Agency (US-EPA) focus on in their guide to indoor air pollution, include Volatile Organic Compounds (VOCs), CO and NO_2 (United States Environmental Protection Agency (2008)). VOCs in indoor air come mostly from products used around the house such as paint, varnish,

cleaning sprays, air fresheners and pesticides but can also be emitted by building materials and furnishings. CO and NO_2 on the other hand are more commonly associated with the use of indoor furnaces, gas cookers, gas heaters, leaking chimneys and people smoking tobacco indoors.

The US-EPA report however, only focuses on indoor air pollution relevant to buildings in the United States (US). In different parts of the world indoor air quality varies in terms of the pollutants and the impact, especially in Asian countries where less clean combustion fuels are often used for cooking in the home and the numbers of people that smoke while indoors is greater (Lee et al. (2010)). Baumgartner et al. (2011) sampled $PM_{2.5}$ in 44 kitchens of the Yunnan area of China in 2010, where 95% of the kitchens used wood or crop residue for cooking, and 96% used a mix of wood-charcoal and wood or crop reside for heating. During the summer months, when the sampling was done, mean concentrations were found to be around 107 μg m^{-3}. Similarly, Li et al. (2011) compared pollutant concentrations in kitchens in relation to different types of stoves in Peru. Means of 181 μg m^{-3} and 3.5 ppm were found for the open-pit stoves for $PM_{2.5}$ and CO respectively. In a larger study across 168 venues in China, Japan, Korea, Malaysia, Pakistan and Sri-Lanka, $PM_{2.5}$ measurements were made and an average indoor level of 137 μg m^{-3} was found (smoking venues were 156 μg m^{-3}, non-smoking venues were 34 μg m^{-3}).

Poor indoor air quality is not always due to indoor sources. The pollutant levels outside an indoor environment have been found to have an impact, although this is dependant on many mitigating factors such as the buildings air filtration units, proximity to outdoor sources, and wind speed/direction. In North America, homes close to Ambassador Bridge (Detroit) were measured over five 24 hour periods Baxter et al. (2008) and it was found that ambient black carbon concentrations significantly contributed to indoor concentrations regardless of wind speed. In Osaka, Japan fine PM ($PM_{2.5}$) was significantly correlated with fine PM outside the properties and it was estimated that about 30% of indoor PM_{10} particles were from diesel exhausts from nearby roads (Funasaka et al. (2000)). In Europe (Prague, Czech Republic) $PM_{2.5}$ was sampled in a school gym during 2005 and 2006, and levels were found to be similar to a nearby fixed-site monitor (24.03 μg m^{-3} compared to 25.47 μg m^{-3}) (Branis et al. (2009)).

2.1.6.2 In-vehicles

The air quality people are exposed to when travelling between indoor micro-environments i.e. a bus or train or car, can differ greatly from general ambient concentrations. Similarly, when space inside vehicles has its own air quality micro-environment. Conditions can be affected by having windows open or closed, the type of vehicle, and the vehicle's location

amongst other factors. Adams et al. (2001a) measured $PM_{2.5}$ during 465 journeys in London over a three week period, at peak and off-peak times of the day during the summer of 1999 and winter of 2000.

Table 2.3: $PM_{2.5}$ by transport mode

Transport mode	**Mean (μg m^{-3})**
Bus	39.0
Car	37.7
Underground tube	247.2
Overground tube	29.3

They observed a great deal of variability between travel modes (see table 2.3), and against typical ambient $PM_{2.5}$ levels (around 10-30 μg m^{-3}) recorded for central London. Outside of London, also in 1999/2000, Gulliver and Briggs (2004) conducted in-vehicle monitoring along a stretch of road in Northampton (80 km North/North West of London). They also observed elevated levels of particulates inside the vehicle (table 2.4).

Table 2.4: In-vehicle PM

PM Fraction	**Mean (μg m^{-3})**
PM_{10}	43.16
$PM_{2.5}$	15.54

Neither of these authors comment on air quality inside vehicles or the resultant exposure. They only conclude that in-vehicle pollutant concentrations cannot be taken to be the same as outdoor values. The situation is further complicated by additional variables such as whether windows are open or closed, the speed of the vehicle, or the number of people inside the vehicle.

2.1.7 Traffic pollution

Between the years of 1950 and 1994, there was a dramatic increase in vehicle traffic on the worlds roads. Vehicle numbers increased from 53 million, to 460 million in the space of 44 years (The World Bank (2013)). Against this backdrop of increasing numbers of vehicles, there is emerging evidence that traffic emissions are harmful to human health. In 2010 the US Health Effects Institute (US-HEI) published the findings of a systematic review of evidence about traffic pollution, and whilst noting that there were many areas still needing further research, there was evidence to support a casual relationship between exposure to traffic-related air pollution and asthma (Health Effects Institute (2010)). Toxicological

research has now also started to link not only primary traffic emissions i.e. exhaust, but also non-exhaust pollutants such as road abrasion, tyre wear and brake wear to adverse health effects (World Health Organization (2013b)). The latter being particularly significant given that there are no laws which consider this element of traffic pollution and therefore no guidelines or limit values. Despite this, measuring and characterising emissions that are solely solely attributable to traffic sources in the urban environment is technically difficult, which makes linkages with exposure estimates and health effects problematic. Different studies have therefore used different pollutant concentrations as markers for traffic emissions. Epidemiological studies have often used NO_2 as a marker for combustion-related pollutants, in particular those emitted by road traffic or indoor combustion sources (World Health Organization (2010)).

Taking the city of Beijing as example again, in 2008 the Olympic Games were held there and this heightened the worlds interest in Beijing's air quality and put the issue under national scrutiny. Global newspapers focused on the effects that poor air quality might have on the performance of the athletes. The reporter James Reynolds of the British Broadcasting Corporation (BBC) wrote "*China is spending billions of pounds on new roads, new venues and on perfect celebratory shows but all that may come to nothing unless this city cleans up its air*" (BBC (2007)). Under this pressure, to try and bring air quality problems under control (at least in the short term while the Olympics were taking place) the Beijing Government implemented a number of measures in the run-up to the games. Stricter vehicle emission standards were adopted, better public transport infrastructure was developed, and from July 2008 to September 2008 a traffic demand management scheme was introduced whereby odd/even vehicle registrations took it in turns to be used on the roads on alternate days (Wang and Xie (2009)). This provided an ideal real-world experiment for the local scientists to attempt to quantify how much poor air quality in the city was attributable to vehicle emissions. Data on black carbon levels was collected by fixed background and fixed road-side monitors, and then analysed to investigate whether the scheme had achieved the desired effect.

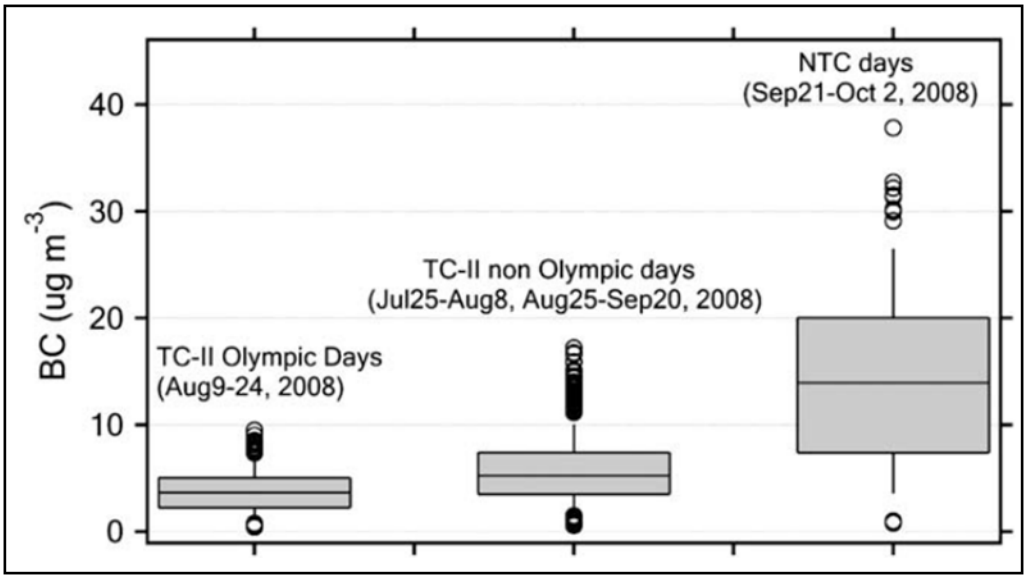


Figure 2.3: Black carbon concentrations during 2008 in Beijing

The results from this study are shown in figure 2.3 and demonstrate how mean black carbon concentrations dropped to around 5 μg m^{-3} during the Olympics (second boxplot) compared to around 14 μg m^{-3} after the Olympics (third boxplot). In addition, during the Olympics on days when the main sporting events were happening, there was a further reduction to around 4 μg m^{-3} (first boxplot). The traffic in Beijing, at least within the limits of this small subset of data, seemed to be contributing to around 10 μg m^{-3} of black carbon pollution in the air. The authors of this study go on to conclude that the main source of emissions in Beijing at the time are from traffic, and that the traffic demand management scheme was effective at bringing emissions down to the (WHO) objective levels. However this seems to be a simplification of the issue, especially given the impact of factories in the vicinity (discussed in 2.1.3). Nonetheless, the exposure of the residents of Beijing to pollution, a debatable proportion of which is from tailpipe emissions, is high.

In Europe, where factories and heavy industry tend not to be based within urban centres, the proportion of the population's exposure to poor air quality, originating from traffic, is high. Often, due to meteorology (see 2.1.4), some emissions may also be from other urban centres. For example emissions from outside London are estimated to account for around 40% of NO_2 concentrations (with the other 60% being generated locally) (Greater London Authority (GLA) (2010). This ratio changes depending on different spatial resolutions. In areas close to roads, the contribution of traffic emissions to overall air quality levels is much higher due to the proximity to the sources (vehicles). The influence of traffic emissions on ambient concentrations is well demonstrated by Mayer (1999). Figure 2.4 identifies clear trends related to rises in CO and NO_2 during morning and evening rush hours on working days.

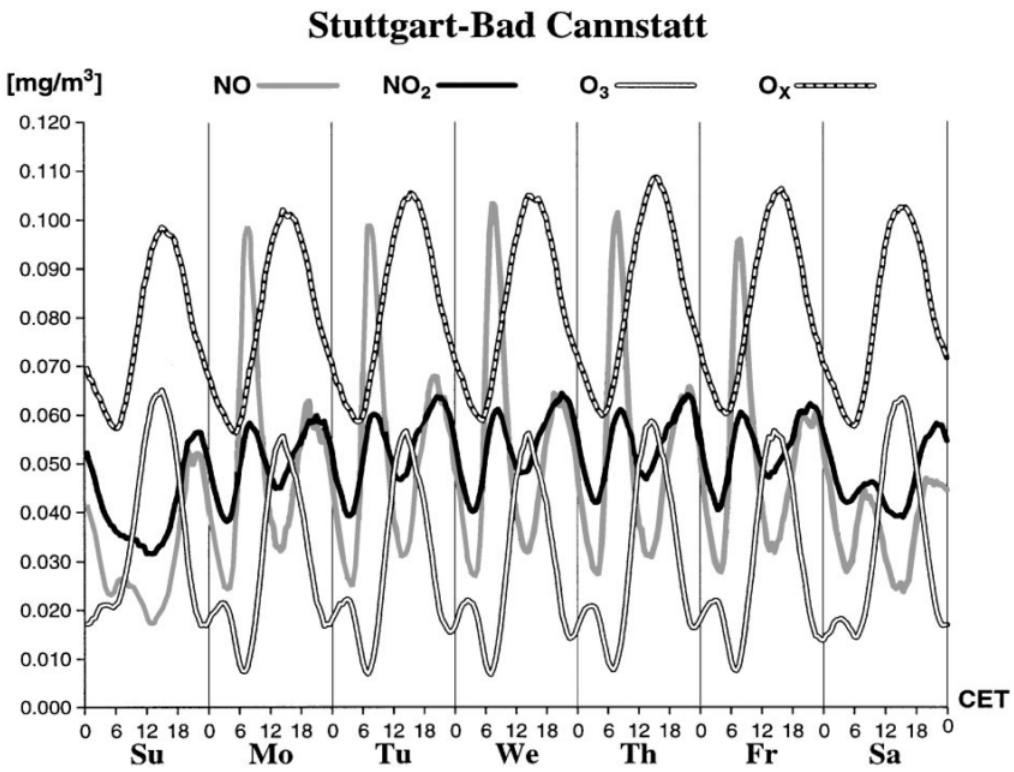


Figure 2.4: Average weekly and diurnal cycles of CO, NO_2 ,O_3 and NO_x at the urban air-quality station Stuttgart-Bad Cannstatta for the period 1981-1993

As air pollution from vehicles is harmful to health, and traffic pollution is so prevalent in urban environments, more permanent long-term attempts to reduce traffic pollution are ongoing in most major cities and countries around the world. Though different countries have sought to achieve this in different ways. An article on 'The Conservation' website Williams (2013)) explains how the EU attempted to legislate to reduce vehicle emissions (Williams (2013)), prompted by the Kyoto Protocol of 1997 which was linked to the United Nations Framework Convention on Climate Change (United Nations (1998)). The protocol sought to specifically reduce CO_2 emissions by a) legislating that car retailers must provide buyers with information on vehicles' fuel consumption and CO_2 emissions and b) placing limits on CO_2 emissions, as a ratio of kilometres travelled, to encourage more efficient use of fuel and therefore lower emissions. The legislation also encouraged companies making vehicles for the EU markets to invest in the production and marketing of diesel vehicles, as they were more fuel efficient than petrol vehicles. In conjunction, these measures were designed to encourage a change in manufacturers behaviour leading to lower emissions through a 'free-market' approach. Unfortunately it became apparent that diesel cars emit higher levels of emissions than petrol cars fitted with three-way catalytic converters (Williams (2013)). The difference between the two emission levels are even greater when considered in the real-world rather than measured in laboratory conditions (Carslaw et al. (2011)). A study

in 2007 estimated that the health effect of favouring diesel vehicles over petrol vehicles in the UK has had the combined effect of contributing to approximately 1850 additional premature deaths over the period 2001-2020, or around 90 premature deaths per year (Mazzi and Dowlatabadi (2007)).

There are of course vehicles that produce low or even zero emissions i.e. hybrid or electric. Hybrid vehicles use a mixture of fuel (petrol or diesel) and a battery, and electric vehicles run solely from a battery. Unfortunately for air quality in urban environments these types of vehicle are currently only a very small percentage of new vehicles, for example in the third quarter of 2018 only 2.2% of new vehicle registrations in London were low or zero emission (for London (2018)).

2.1.8 Summary

Section 2.1 introduced the subject of air pollution – a non-naturally occurring material in the air, altering it's composition. It can take many different forms and be categorised in different ways, for example particulate matter, nitrogen oxides, ozone or sulphur dioxide. It was discussed how many of the causes of air pollution are linked to vehicle combustion engines and that in urban environments, where people are increasingly living, this places the sources and public in close proximity to each other. This can be affected (both positively and negatively) by different meteorological conditions and the topology of the region, city, and even individual streets and buildings. Within these environments, it was explained that there are micro-environments such as inside vehicles and buildings which can also raise or lower pollution levels. As this research intends to focus on urban environments, traffic emissions were then considered in a little more detail. The Beijing Olympics 2008 was used as a case-study to understand the impact that traffic emissions can have on air quality in a major city, and the diesel dominated vehicle fleet of Europe was then explained (touching on the impact compared to petrol that this has had on air quality).

Now that the subject of air pollution has been introduced, the next section of this background will give an overview of the impact on human health from air pollution i.e. why Section 2.1 actually matters to us.

2.2 Health effects of air pollution

> "Clean air is considered to be a basic requirement of human health and well-being. However, air pollution continues to pose a significant threat to health worldwide" (World Health Organisation (2006)).

2.2.1 An overview

From the 1600s onwards coal was the main source of heat and energy in major UK cities. Concern from the scientific community and coherent programs of research about the possible negative effects of this fuel source were limited. When undertaken the research often focused on poor visibility or damage to buildings rather than human health. It was the early 1900s when coherent and robust studies began to investigate mortality and links to fog, as it was known at the time. Notably with Russells paper *'The Influence of fog on mortality from respiratory diseases'* being published in The Lancet in 1926 (Russell (1926)). This publication preceded London's 'Great Smog' of 1952, which was one of the UK's most important air pollution events in history in terms of realisation of the links between pollution and health. Research conducted since this event has had a great impact on the study of air pollution, public perception and government regulation to combat it. Data at the time showed that a rise in fog (pollution) levels was closely followed by rises in mortality and morbidity (Bell et al. (2003)). At the time it was estimated that between 3,500 and 4,000 more people died than would have normally been the case in this period (See figure 2.5 from Greater London Authority (GLA) (2002)). The rise in mortality was originally attributed to influenza, however sensitivity analysis by Bell et al. (2003) revealed that only an extremely severe influenza epidemic could have accounted for the excessive deaths recorded for that period. Subsequent reanalysis of the data estimated that between December 1952 and March 1953 there were actually 13,500 more deaths than during the same time period the previous year, attributable to (controlling for temperature and influenza) rather than the 3000–4000 generally reported for the episode.

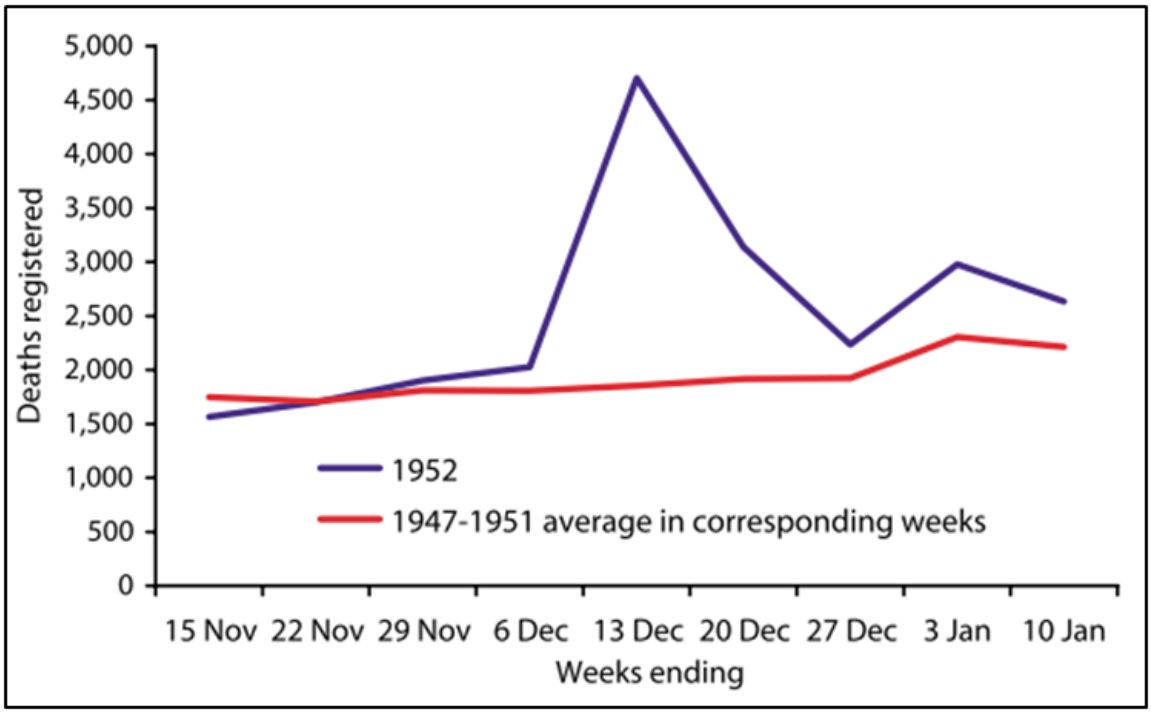


Figure 2.5: Recorded deaths comparison during 'Great smog' period. 1952 (blue line) shows a peak in deaths coinciding with the 'great smog' which is not seen in the preceding or following years (red lines)

Once this explicit link between air pollution and health became apparent, laws and regulations began to be written and passed. For example the Clean Air Act in 1956 (with various revisions over time, notably in 1968), the 1970 European Commission (EC) Directive (70/220/EEC), the 1974 Control of Pollution Act and the 1979 International Convention on Long Range Transboundary Pollution. It is now widely accepted that air pollution has harmful effects on human health (World Health Organization (2013b). Although in the Western world, the sources of pollution have shifted from using coal for heating and cooking, to be dominated by combustion engines in vehicles and similar (as discussed in Section 2.1.7).

When considering the effects of air pollution on public health, studies on large groups of people (tens of thousands plus) often use epidemiological methods. Studies using epidemiological methods will be discussed many times during this thesis, therefore a brief definition of epidemiology and the key terms are now given.

Epidemiology is the the study of how often disease/poor health occurs in a group of people, and the factors that lead to it. Or, more technically defined by Bonita in a WHO publication as *'the study of the distribution and determinants of health-related states or events in specified populations, and the application of this study to the prevention and control of health problems'* (Bonita et al. (2006)). Some key terms include (adapted from U.S. Department

of Health & Human Services (2014)):

- Incidence: The number of new ill people in the population over a specified time period
- Prevalence: The existing number of ill people in the population over a specified time period.
- Burden of disease: The total significance of the disease or illness to wider society. For mortality this is often measured in years of life lost.
- DALY (Disability-Adjusted Life Year): A statistic to represent the health of a population. One DALY represents one lost year of healthy life and is used to estimate the gap between the current health of a population and an ideal situation in which everyone in that population would live into old age in full health.

Epidemiological studies have '*For decades [...] been a cornerstone of our approach to investigating the health effects of air pollution and have been a principal basis for setting regulations to protect the public against adverse health effects*' Zeger et al. (2000). Recent high profile examples that look at air pollution include, but are not limited to; respiratory problems (Peacock et al. (2011)), cardiovascular issues (Brook et al. (2010)) and cancer (Pope III et al. (2012), Loomis et al. (2013)). Indeed, recently (17 October 2013), the International Agency for Research on Cancer (IARC) classified outdoor air pollution as carcinogenic to humans (Loomis et al. (2013)). In an IARC press-release, Dr Kurt Straif, Head of the Monographs Section stated *"The air we breathe has become polluted with a mixture of cancer-causing substances. We now know that outdoor air pollution is not only a major risk to health in general, but also a leading environmental cause of cancer deaths"*.

However as discussed in Section 2.1.2, the term 'air pollution' covers many different pollutants. It is important therefore to untangle which pollutants are more or less harmful to health. This will help people avoid areas with pollution that is most harmful, and help politicians to develop policies that are effective at reducing the most harmful types of pollution. For example there is good evidence that the $PM_{2.5}$ shortens life, however it tends to be emitted in locations where there are sources of other pollutants such as NO_2, which makes it hard to disentangle the effects of them individually (Committee on the Medical Effects of Air Pollutants (2018)). A broad overview of epidemiological studies on the health effects of exposure to $PM_{2.5}$ is now discussed.

Worldwide, WHO estimate that $PM_{2.5}$ causes about 9% of lung cancer deaths, 5% of cardiopulmonary deaths, and 1% of respiratory infection deaths (World Health Organization (2012)). In 2013 the Global Burden of Disease publication ranked exposure to air pollution and particulate matter as one of the top ten risk factors for health globally, estimating

that over 430,000 premature deaths and around 7 million years of healthy life were lost in Western, Central and Eastern Europe in 2010 from exposure to fine particulate matter (Brauer et al. (2012)).

In a study looking at $PM_{2.5}$ and anthropogenic ozone, Silva et al. (2013) recently modelled ozone and $PM_{2.5}$ surface concentrations for the entire world, then used concentration-response functions for long-term exposure and mortality (from an American Cancer Society publication) to estimate that annually and globally there are about 2.1 million premature deaths from respiratory problems linked to $PM_{2.5}$, with these being split 93:7 between cardiopulmonary disease and lung cancer (Silva et al. (2013)).

We can therefore see that the health effects of $PM_{2.5}$, when taken in context of large populations are significant. However, these figures are likely to be the tip of a much larger concern as most do not include morbidity i.e. poor health and detriments to the populations quality of life that do not result in death. This is illustrated by figure 2.6 (Mannino (2000)) showing a greater number of the population have less severe health effects that still have a burden on public well-being.

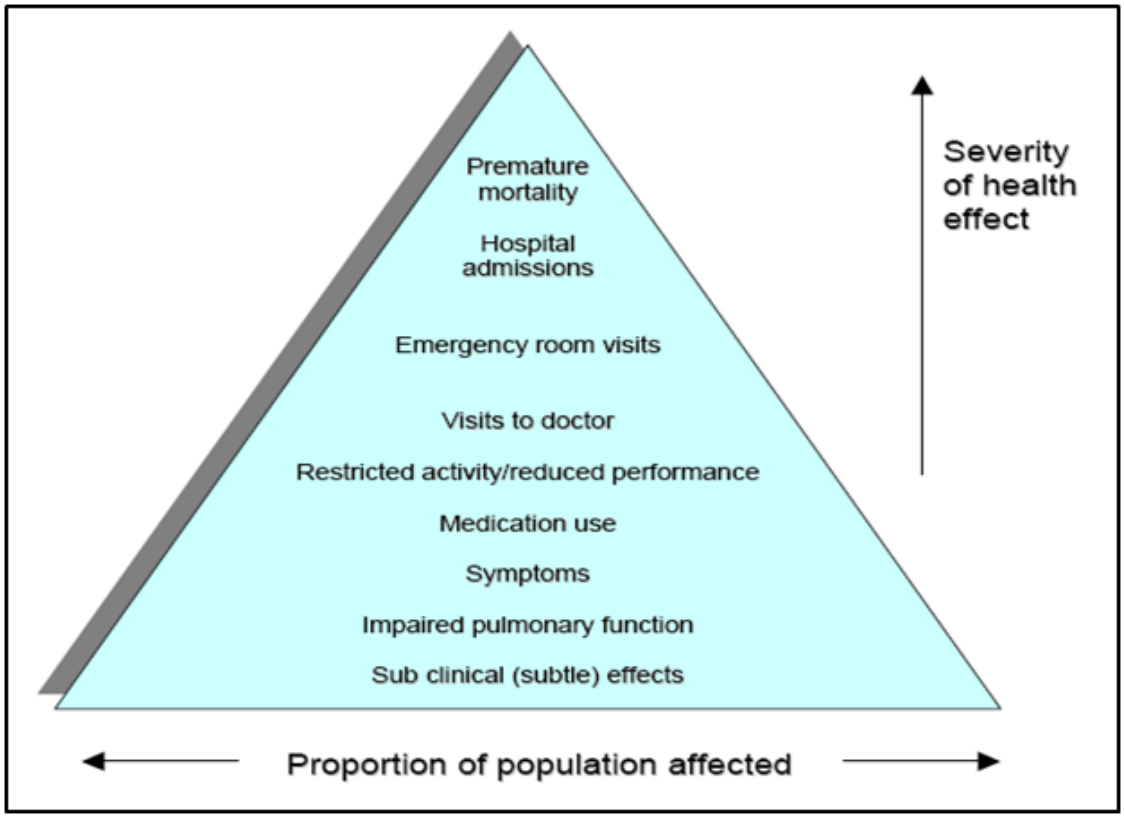


Figure 2.6: Air pollution health effects pyramid

In the UK, the Committee on Medical Effects of Air Pollutants (COMEAP) has been set-up to provide advice to the government and related agencies via the Department of Health's

Chief Medical Officer on the harmful effects of air pollution. COMEAP regularly publish reports summarising findings into the health effects of pollutants. In their 2010 report on mortality, they estimated that around 29,000 deaths in the UK in 2008 were attributable to $PM_{2.5}$ (Committee on the Medical Effects of Air Pollutants (2009)). Expressing this differently, they estimated that air pollution may have contributed to the earlier deaths of about 200,000 people in 2008, with an average loss of life of about two years per death affected. Also on a UK scale, Yim and Barrett (2012) estimated that $PM_{2.5}$ causes about 19,000 premature deaths in the UK and 3,200 in London. Note that the reason for these figures being lower than that of COMEAP is that this study focused solely on the deaths attributable to traffic emissions rather than all sources (and gives at least in this context a rough idea of the impact of traffic emissions compared to non-traffic emissions). At the city level, Miller (2010) used life-tables and concentration-response coefficients to estimate deaths attributable to $PM_{2.5}$ to be 4,267 in Greater London for 2008.

While epidemiological studies find strong links between air pollution and poor health, particularly pulmonary and cardiovascular disease, the mechanism of **how** air pollution causes mortality is not yet fully understood.

Chamber studies are often used to identify the toxicological effects of air pollutants. Human subjects will be medically examined before entry to the chamber, then sealed inside for a set period of time, and re-examined after exposure to pollutants. The atmosphere within can be carefully controlled by investigators to simulate the environment of their choice, in the case of air pollution normally this is heavily polluted air. Lung function tests, haematology and fiberoptic bronchoscopy are undertaken to investigate the pathological pathways by which pollutants may cause disease and poor health.

Salvi et al. (1999) exposed 15 healthy human volunteers in an environmental chamber to clean air and then diluted diesel exhaust, for one hour at a time. Significant increases in neutrophils and platelets, markers of stressed airways and lungs, were observed in the subjects peripheral blood after the diesel exposure, however lung function measured before and after exposure revealed no decline. The study demonstrated that at high concentrations of diesel exhaust, at least in the short term, there is a systemic and pulmonary inflammatory response in healthy human volunteers, which is not detected by standard lung function measurements alone. A further study by Salvi et al. (2000) exposed healthy human volunteers in chambers to a range of particles/diesel exhaust and fiberoptic bronchoscopy was performed six hours after each exposure, the results suggest airway leukocyte infiltration as the underlying mechanism for diesel exhaust-induced respiratory health outcomes. In a study by Ghio et al. (2000) no immediate symptoms were observed, however 18 hours after exposure inflammation was identified in the lower respiratory tract, particularly in those with

the highest particulate exposure compared to clean filtered air.

The afore-mentioned studies suggest that pulmonary inflammation in the airways is responsible for damage to the lungs. Although more studies are needed to pin-point the specific pathways. Chamber studies have helped to understand the inflammation pathways following short-term, however further issues also need to be addressed, for example, the lack of clarity between the length of exposures and the subsequent responses and how long the lag is between exposure and underlying biological mechanism responses (such as in the studies just mentioned where different effects were noted after 6 hour and 18 hours) (Environmental Protection Agency (2009)).

To summarise, the American Heart Association (AHA) Brook et al. (2010) describe the probable mechanism between particulate matter air pollution and cardiovascular disease as the following:

1. Release of proinflammatory mediators (eg, cytokines) from activated immune cells, or platelets or vasculoactive molecules (eg, Endothelin, histamine), or microparticles endothelium of blood vessels in the lung
2. Perturbation of systemic autonomic nervous system balance or heart rhythm by particle interactions with lung receptors or nerves
3. Translocation of PM (ie, ultrafine particles) or particle constituents (organic compounds, metals) into the systemic circulation

A simplified diagram from the same reference (Brook et al. (2010)) has been adapted and is shown below (fig 2.7) for illustrative purposes.

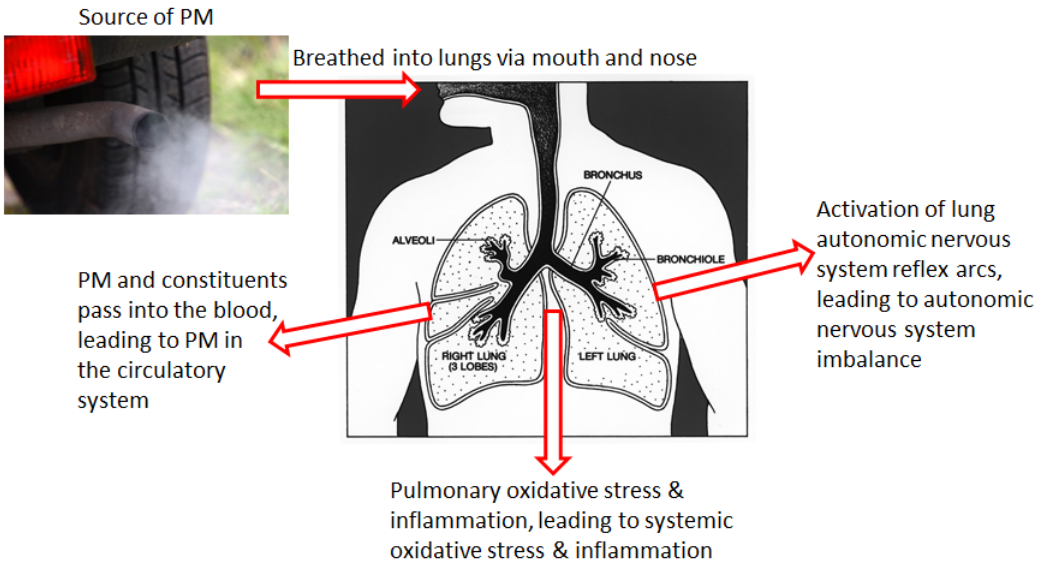


Figure 2.7: The biological pathways linking PM exposure with cardio–vascular disease

A detailed review of the biological mechanisms underlying disease is not the scope of this thesis. Rather it is concerned with the methods of estimating levels of exposure to air pollution. The presumption of this research is that even low levels of exposure to air pollution is harmful. This is justified by studies such as the European ESCAPE project (Beelen et al. (2013)) which found that long-term exposure to fine particulate pollution is associated with mortality, even at concentration ranges well below the present European annual mean limit values. The work of Brauer et al. (2002) and his group looked at the use of minimum threshold values for exposure to $PM_{2.5}$ and found that due to exposure miss-classification, population-level thresholds were apparent at lower ambient concentrations than common personal thresholds (such as the EU limit values discussed in table 2.2 of Section 2.1.3).

In summary, epidemiological and toxicological studies have shown air pollution is a major environmental risk to health. By reducing air pollution levels, and exposure to air pollution, governments should be able to reduce respiratory symptoms, heart disease, and lung cancer in their population. In order to calculate the estimates of the detrimental effects on people's health, exposure and dose must be calculated as accurately as possible. That is the amount and types of air pollution that people are exposed to, and breathe in,every day of their lifetime. This can be done in many different ways, and methods are undergoing regular revision as scientists attempt to improve their accuracy.

One of the key issues to address to further understand the links between pollutant exposure and poor health is the duration of exposure. Does long-term exposure to low levels of pollution have the same effect as short-term exposure to high levels of pollution in terms of disease prevalence or DALY The importance of short-term versus long-term exposure on health is explored in the next section.

2.2.2 Long term exposure v. short term exposure

There are few studies that compare the effects of long and short-term exposure to air pollution on a large population, at adequate spatial and temporal scales, for a range of pollutants, and with appropriate health information to enable us to determine which has the most impact on health.

This section therefore presents a selection of the short term and long term studies conducted to date, as well as the small number of studies that have attempted to reconcile the two. This is an issue that is being wrestled with by the community at present.

The most widely cited large-scale long-term exposure/health studies are on the effects of NO_2 in New Zealand Scoggins et al. (2004), fine particles globally in the Global Burden

of Disease 2010 (Brauer et al. (2012)) and a meta-analysis review of both by Faustini et al. (2014). Scoggins modelled annual average NO_2 concentrations over 3 km x 3 km grid squares in Auckland for the years 1996-1999 and linked these to mortality data for the same region provided by the Health Information Service (estimating a 1.3% increase in mortality for each 1 μg m^{-3} rise in NO_2 annual average values), Brauer used global estimates of ozone and $PM_{2.5}$ concentrations at a 0.1° x 0.1° spatial resolution to inform the global burden of disease modelling (discussed previously in Section 2.2.1), and Faustini's review brought together studies that looked at NO_2 and $PM_{2.5}$ on population mortality and found that there was rise of 1.04% and 1.05% per years of life per 10 μg m^{-3} for NO_2 and $PM_{2.5}$, respectively. Meta-analysis by Hoek et al. (2013) estimates an excess risk of 6% per 10 μg m^{-3} increase in $PM_{2.5}$ exposure.

As stated, links between long-term exposure and poor health seem to be fairly robust, however as long exposure time-frames that are used (typically annual average concentrations), and are combined with large spatial scales of hundreds of kilometres, it is possible that much of the detail in the results is being missed - which will be discussed more in sections 2.3 and 2.4. Furthermore, no short-term exposure on health effect outcomes from poor air quality are considered in these 'long-term' studies, or a comparison of the importance of long-term versus short-term exposure effects.

On the flip-side, short-term epidemiological exposure health studies tend to suffer from similar issues from the perspective of someone trying to weigh-up short-term versus long-term exposure. The review by Brook et al. (2010) (briefly discussed earlier in 2.2.1) concluded that short-term epidemiological time-series studies tend to find that around a 10 μg m^{-3} increase in mean 24-hour $PM_{2.5}$ concentrations increases the risk of cardiovascular mortality by approximately 0.4% to 1.0%. But that importantly this risk is not distributed evenly across the population i.e. people with existing medical conditions or the elderly are more vulnerable to effects triggered by this short-term exposure (which comes back to the point in the previous paragraph about missing detail). Furthermore, a review of time series studies by Atkinson et al. (2014), looking at relationships between $PM_{2.5}$, daily mortality and hospital admissions found similar percentage increases to Brook. A 10 μg m^{-3} in $PM_{2.5}$ was associated with a 1.04% increase in mortality. However a comparison in the same datasets with long-term exposure is not available.

Two recent studies that have sought to answer the question of short-term versus long-term exposure, or at least explore it, are those by Kloog et al. (2013) and Beverland et al. (2012b). Kloog et al. geo-coded all deaths in Massachusetts (USA) between the years 2000-2008, modelled short-term and then long-term $PM_{2.5}$ concentrations for the area of their data-set, and then used time-series analysis to try and examine the relationships. They found for

short-term relationships (day of death, and three days prior) that every 10 μg m^{-3} in $PM_{2.5}$ there was a 2.8% increase in PM mortality. Then for long-term exposure they found that for every 10 μg m^{-3} increase in $PM_{2.5}$ the odds of death occurring rose to an odds-ratio of 1.6 (which simplistically equates to around a 60% increase). Leading them to conclude that the effects of long-term air quality appear much more pronounced than short-term. Contention remains whether their definition of short-term is the most appropriate. Since the investigators only looked at $PM_{2.5}$, it might be that the effects of $PM_{2.5}$ are more pronounced in the longer–term, but NO_x may have the opposite relationship. More data collection and analysis is needed.

Beverland et al. (2012b) take a similar approach to Kloog et al. by retrospectively using mortality data from the 'Renfrew–Paisley' and 'Collaborative' cohort studies which began in the 1970s in Glasgow (Scotland), linked to black-smoke data (the only pollutant routinely measured at the time). They too considered short-term to be three days lag, and found that there were short-term exposure-mortality associations greater than those found in the general population, indeed rises in black smoke levels were affecting mortality. Furthermore, there were also long-term mortality associations, which were more strongly associated than the short-term associations. They conclude, similarly to Kloog et al., that long-term associations have more impact than short-term. As this study was carried out retrospectively the collection of data and population definitions were not ideal. In particular, the way that the black smoke exposure was estimated using monitoring stations for short term, but modelled for long-term, introduces uncertainty. Also for comparing their associations, they used the general population for short-term and the cohort for long-term, therefore differences in these groups could cause bias in their results.

COMEAP published a report in 2010 titled *'The Mortality Effects of Long-Term Exposure to Particulate Air Pollution in the United Kingdom'*. From reviewing evidence of published studies they too concluded that long-term exposure was a more important factor in contributing to mortality than short-term exposure (Committee on the Medical Effects of Air Pollutants (2010)). Finally, bringing together many data-sets and conclusions from other studies, Stieb et al. (2002) produced a number of forest plots for different pollutants demonstrating the effect of short-term exposure. The NO_x plot is shown in figure 2.8 which shows a consistent finding of an increase in mortality as NO_2 increases.

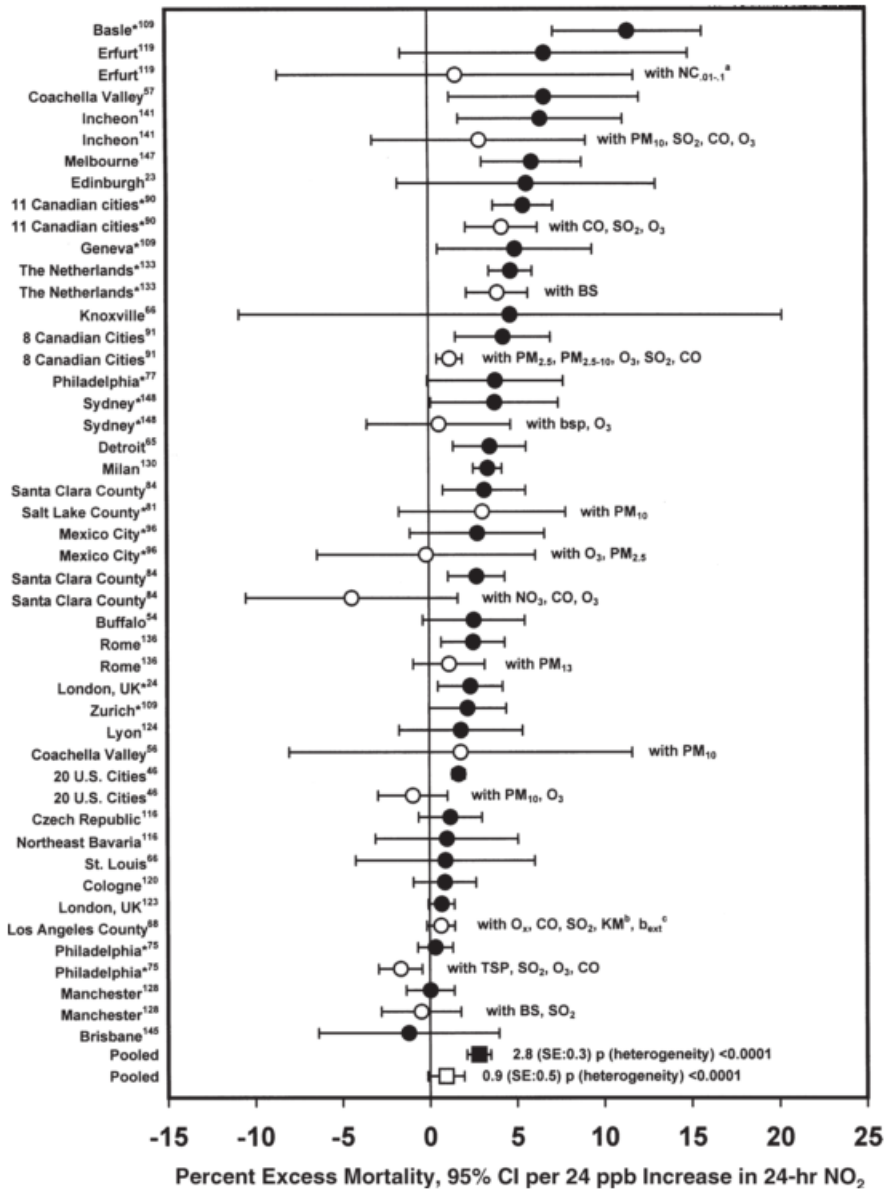


Figure 2.8: Percent excess mortality

To summarise, the research that has been conducted so far, with the data available, suggests that the effects of long-term pollution are a more important determinant of poor health than that of short-term exposure. However the studies are limited and lack detail sometimes in areas which potentially alter their conclusions. Kloog et al. and COMEAP looked at $PM_{2.5}$ (though perhaps this is sensible given the discussion in Section 2.2.1), and Beverland et al. only considered black smoke. There were limitations of spatial and temporal resolution of the pollution data and subsequent linkage in the three of the studies. Often large data-sets cannot consider all important details. Definitions of how long 'long-term' and how short 'short-term' are could also be an influence on conclusions drawn. Short-term is taken to be around 3 days exposure, but this could be missing hyper-short term exposure, such

as a subject spending an hour cycling down a busy road and perhaps triggering hospital admission for breathing difficulties. With better air quality data and more in-depth better information on where people spend their time (and therefore are exposed) it might be possible to overcome these issues, however the data to do this does not presently exist for epidemiologists to use.

It is important to consider both the rapid effects of air pollution exposure e.g. pathways without hours of exposure **and** the chronic effects of sustained exposure (Brook et al. (2010)).

It seems that short-term (days, months), hyper-short-term (hours, minutes) and long-term exposure (months, years) are all important in understanding the health effects of air pollution. Thus studies that are able to consider large population exposures (but with information on individuals available), and on these varying time–scales are required. This issue of improving linkage between air quality and exposure has evolved over time. The next section reviews the various methods that have been used in the past to link air quality and exposure time – henceforth termed 'static exposure studies' i.e. studies that make use of various different metrics for air quality, but do not take into account the movement of people. Dynamic exposure studies are then described – those that take this movement and other factors into account.

2.3 Static exposure & health studies

The methods of estimating individual and population level exposure to air pollutants (and from this data the impact on health) has evolved over time; better data has become available, computer modelling has become more complex (facilitated by more powerful computers), and more accurate methods have been employed. This section of the PhD takes an overview of 'static exposure' studies. That is, the subjects to which the air pollution is being attributed do not move between environments and their exposure is calculated (normally) at their residential address. In addition the temporal and spatial granularity of the air quality data is often of low or insufficient quality.

This section is split into the following categories:

- Large area exposure
- Monitoring stations
- Proximity to roads
- Dispersion modelling
- Land-use regression

Due to a large literature base, these sections draw on key texts for each section as examples of the approach being described. Having considered these studies, the use of dynamic exposure models are then introduced in Section 2.4, ('Dynamic exposure & health studies').

2.3.1 Large area exposure

Large area exposure studies are studies which, as the title suggests, attribute exposure to subjects over a large spatial area. This might be at the level of a city, a county or even continent. The well known Global Burden of Disease studies (discussed briefly in 2.2.2) are classic examples of this approach. For the global burden of $PM_{2.5}$ (chosen due to it's strong links in the literature with poor health) in 1990 and 2005 an annual average layer of air pollution was modelled for the entire world at 0.1° x 0.1° spatial resolution (shown in figure 2.9).

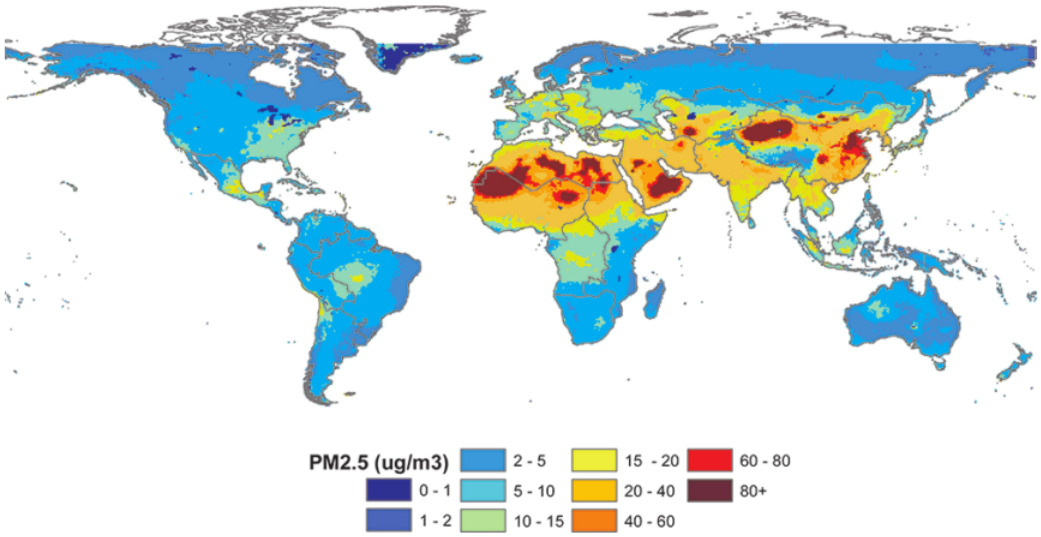


Figure 2.9: Estimated 2005 annual average $PM_{2.5}$ concentrations (ug/m^3)

To generate this worldwide layer of $PM_{2.5}$ satellite derived observations of Aerosol Optical Depth (AOD) were used (Brauer et al. (2012)). When this work was undertaken this was not a particularly common approach for exposure assessment, mainly due to lack of understanding in this new field and the resolution being relatively low, however as new earth observation satellites are launched the approach is gaining in popularity (van Donkelaar et al. (2015), Hoek (2017)).

The resulting health conclusions/estimations from large scale exposure models are discussed in Section 2.2.2, here we focus on limitations with the methods. Firstly, ambient concentrations are assigned to the entire population of the grid cells i.e. no micro-environmental modelling is undertaken to allow for the time that people spend indoors, or indeed in any other environment. Secondly, modelling at 0.1° x 0.1° resolution means much of the spatial variation in concentrations is lost i.e. there is a great deal of small scale variation in concentrations within the area which are not being included. To elucidate further, we know from Section 2.1.3 that people are concentrated in urban environments rather than distributed evenly across countries and continents, and that within these environments levels of pollution are higher than outside of them, primarily due to emissions from combustion engines (Section 2.1.7). Taking London as an example, the city is approximately 50 km wide, yet a degree of longitude is approximately 113 km wide, meaning that (in this oversimplification example) the exposure attributed to someone living in the middle of the Kent downs where are much fewer sources of $PM_{2.5}$ is the same as someone living in the centre of London, where the sources of $PM_{2.5}$ are more frequent. The authors argue that as the population data is of a similar scale this does not matter so much, however by not being able to take account of these urban environments, or at least not at an adequate scale, exposure may

be incorrectly attributed. Though to what degree it is hard to know. Thirdly, the lack of temporal resolution to the air quality layer is a problem. As was briefly explored in the sections on traffic-generated pollution (2.1.7) and meteorology (2.1.4), air quality varies by hour, week, days, months, seasons etc.

Studies of similar scales include Silva et al. (2013) who combined 14 atmospheric chemistry and meteorological models to attribute annual average $PM_{2.5}$ exposure to the worlds population on a scale of 0.5° by 0.5° degrees resolution, and Boldo et al. (2011) who modelled $PM_{2.5}$ at a resolution of 18 km by 18 km (figure 2.10) to cover the country of Spain (both approaches then used concentration-response functions to estimate mortality).

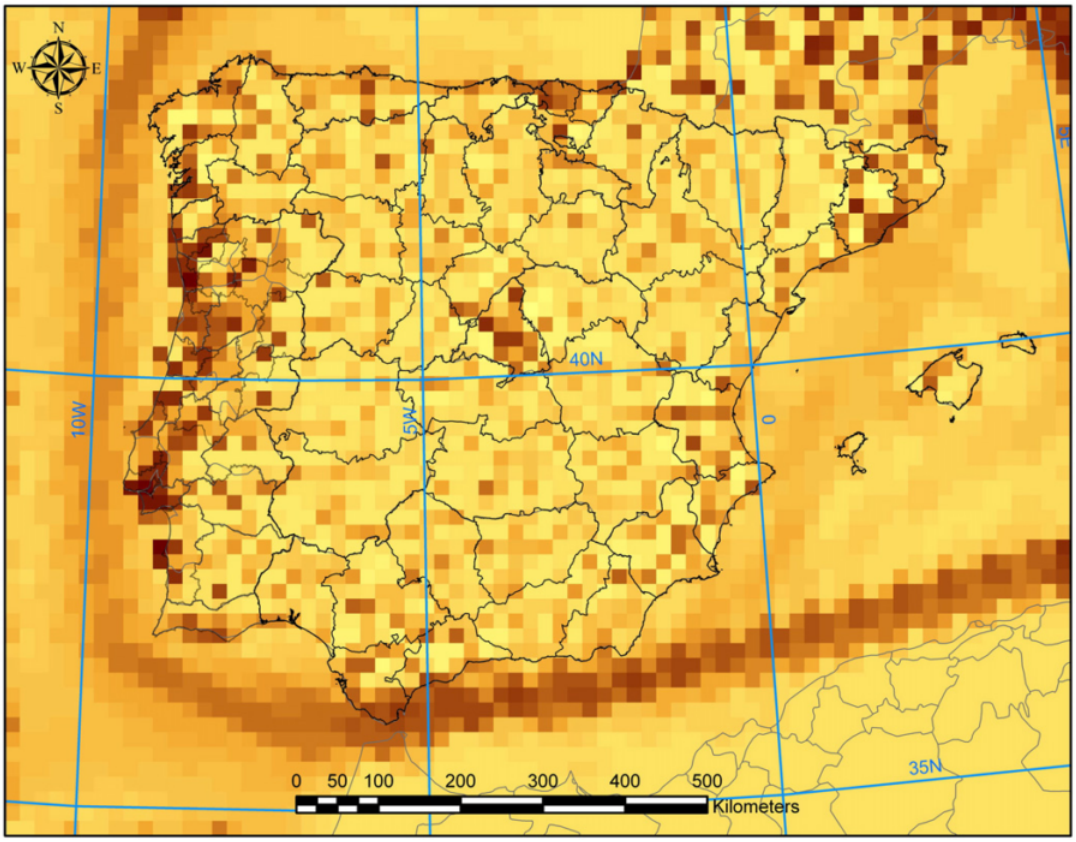


Figure 2.10: Grid squares used for $PM_{2.5}$ exposure

It should be stated at this point, that the criticism levelled at these papers is symptomatic of many studies discussed in the next sections of this report. However doing so is often a little unfair. The researchers were in most cases doing the best work that technology, available data, or indeed time or known methods allowed them to do. As the discussion moves into reviewing further literature in Section 2.4 (Dynamic exposure & health studies), the suggestion is that methods can now be improved beyond this position.

2.3.2 Monitoring stations

The term 'air quality monitoring station' or simply 'monitoring station', within the context of this field of research, typically refers to a static cabin or large metal box which houses a number of instruments that measure various air pollutants and meteorological conditions at specific time intervals. A typical station is shown in Figure 2.11 from `www.londonair.org.uk`.

Figure 2.11: A typical monitoring station

In the UK monitoring stations were first set-up after the introduction of the Clean Air Act in 1956, and the numbers of stations, locations, the accuracy, time resolution and numbers of pollutants that are monitored has changed incrementally since then. The last change in organisation of the sites came in 1998, when the then Department of Environment established the Automatic Urban and Rural Network (AURN), bringing together many sub-networks to form the most comprehensive automatic national monitoring network in the country, made up of 127 sites (DEFRA (2011b)). The data collected from the sites is used for policy and legislative purposes, including reporting to the EU (long may this continue), however as the data is routinely and automatically collected over long periods of time the outputs have been a popular dataset for researchers too. Mayer (1999) presents a good example of the temporal variability of data that can be collected such as CO, NO_2, O_3 and NO_x at an urban air quality station in Stuttgart, Germany (and then subsequently linked to populations for estimating exposure).

One of the most highly cited research articles where data of this nature was used is in the paper '*An association between air pollution and mortality in six US cities*' by Dockery et al. (1993). In this study air pollution exposure was linked to 8111 adults over six cities in theUS from six monitoring sites (one in each city) in the form of annual average concentrations. Each station measured $PM_{2.5}$ and PM_{10} (PM_{15} before 1984). After adjusting for smoking and other risk factors, significant associations between air pollution and mortality were found. Studies that also used monitoring stations in a similar way include Atkinson

et al. (2010) who did a time-series analysis of London hospital admissions using data from a background monitoring site (PM_{10}, $PM_{2.5}$, $PM_{10-2.5}$) and found associations between various pollutants and health outcomes outcomes such as daily mortality (particular for cardiovascular), and then Samoli et al. (2004) who as part of the APHEA Multi-city Project took monitoring site data for 22 EU cities and found links with mortality from the air quality data of these stations.

The main question with this type of approach however is that of temporal and spatial variability, and resolution - can the data from a monitoring site be an appropriate measure of exposure for someone who, in many examples, may live miles from the monitoring site. Certainly the studies that have tried to address whether this is appropriate or not have found poor correlation between the two. Cyrys et al. (2008) found that '*the use of a single monitoring station in long-term epidemiological studies must be insufficient to attribute accurate exposure levels of PNCs to all study subjects*' i.e. it might work for some people but not for all. Although it's worth noting here that they believe that the monitoring stations do an accurate job of reflecting temporal variability, particularly for ultra-fine particles. Just not the spatial variability. Goldstein and Landovitz (1977) broadly agree, they found in New York that '*the procedure of using one aerometric station to represent the daily fluctuations of air pollution throughout the large metropolitan area of New York City risks the use of an unreliable or invalid measure of the short term variation in air pollution'*. So their message is slightly different, in that their research did not assesses how valid it would be as a measure of long-term variation, but they did conclude it is poor for short-term (whether long-term or short-term exposure to poor air quality is more important as a measure is discussed in Section 2.2.2, Long term exposure v. short term exposure). Willocks et al. (2012) in Scotland extended two existing studies of the relationships between air quality at monitoring sites and cardiovascular disease, and more broadly reflected on the difficulties in conducting this sort of study, and found *"no consistent associations [...] between PM_{10} concentrations and cardiovascular hospital admissions"*.

Whether this mis-classification of exposure to pollutants varies between pollutants and height from the ground was considered by Restrepo et al. (2004) who took data from three different monitoring stations (15m above ground) and compared it to data from a van which contained similar equipment but which was parked at three different locations and with the equipment inlets at 4m above ground.The stations showed good agreement between themselves, but not with the ground-level (van) data. $PM_{2.5}$ was closest matched, for ozone the ground level concentrations were generally lower, and for NO_2 the concentrations at ground level were over twice as high as those at the monitoring stations.

To conclude, using data from monitoring stations as a proxy for estimating exposure for large populations over many kilometres does not seem to accurately reflect individuals exposure. The studies above tended to look at the correlation between a subjects residential address and the monitoring station, meaning that there is then the additional complicating factor that people do not spend all of their time at their home. Whilst the data is certainly easily obtained and the temporal resolution makes it attractive for time-series studies, the lack of spatial variability and lack of any micro-environmental modelling or understanding of where subjects actually spend their time mean that this approach is simplistic.

2.3.3 Proximity to roads

The use of subject's address data, and then a calculation of the number of roads (and sometimes traffic density on those roads), is another proxy measure of exposure that has been used to consider exposure to air quality and links between this and poor health. One of the most highly-cited papers in this area is by Gauderman et al. (2007) who specifically looked at whether living near to major roadways in California had an impact on lung-function growth of children between the ages of 10 and 18. In the study 3677 children were regularly monitored for 8 years and yearly lung-function tests were completed, which were then considered alongside their home address and the distance to the nearest freeway. They found that proximity to freeway traffic is associated with substantial deficits in lung function (which became less pronounced the further the child lived from the freeway). Similar studies include those by Janssen et al. (2001) and Rose et al. (2009), although both do not go as far as to make conclusions of health outcomes, they focus on the method of exposure estimation by road density. A wider-review of studies in this area was also undertaken by the Health Effects Institute (2010) in the report '*Traffic-related air pollution: a critical review of the literature on emissions, exposure, and health effects*'.

The presumption of this type of study is that the air quality the subjects are exposed to during normal day-to-day activities is strongly correlated to the air quality at their home, and that the air quality at their home is strongly correlated with the number of freeways within certain buffer distances of their home. These studies also mostly look at annual average traffic flows or similar, and therefore seem to not take account of variation in road use and therefore pollution levels. Indeed, combining these issues only exasperates the uncertainty, for example road flows are mostly higher during the morning and evening rush-hours, when children between 10 and 18 are likely on their way to school or already at school i.e. not at home as the model presumes. They also don't consider the time a subject is not at their home, or any other micro-environment.

2.3.4 Dispersion modelling

To better bridge this distance between where exposure is occurring (often presumed to be the subject's household) and where the exposure is being estimated or measured (such as a monitoring site) modelling can be used to create maps or layers of varying resolution for the area that the study or subjects are based in.

Dispersion modelling is a way of simulating the movement of emissions through the atmosphere using mathematical equations with input variables such as wind speed, wind direction, air temperature and street geography (Environmental Protection Agency (2008)). By doing this at different scales (country-wide, city-wide, individual streets) it is possible to create pollution maps and to then associate the pollutant values at those places with the people who live there - estimating their exposure - and possibly linking this to health effects if data is available.

Maroko (2012), in studying environmental justice in New York City (USA) compared the differences between using a proximity analysis technique (similar to 2.3.3 - proximity to roads) and a dispersion model of $PM_{2.5}$. Their hypothesis was that minority populations were more likely to be located in areas of poor air quality, and that proximity analysis may under-represent this problem. Figure 2.12 shows tax-lots within $1/4$ mile of the point stacks upon which the initial exposure analysis (and subsequent assessment of over-representation of minorities) was completed.

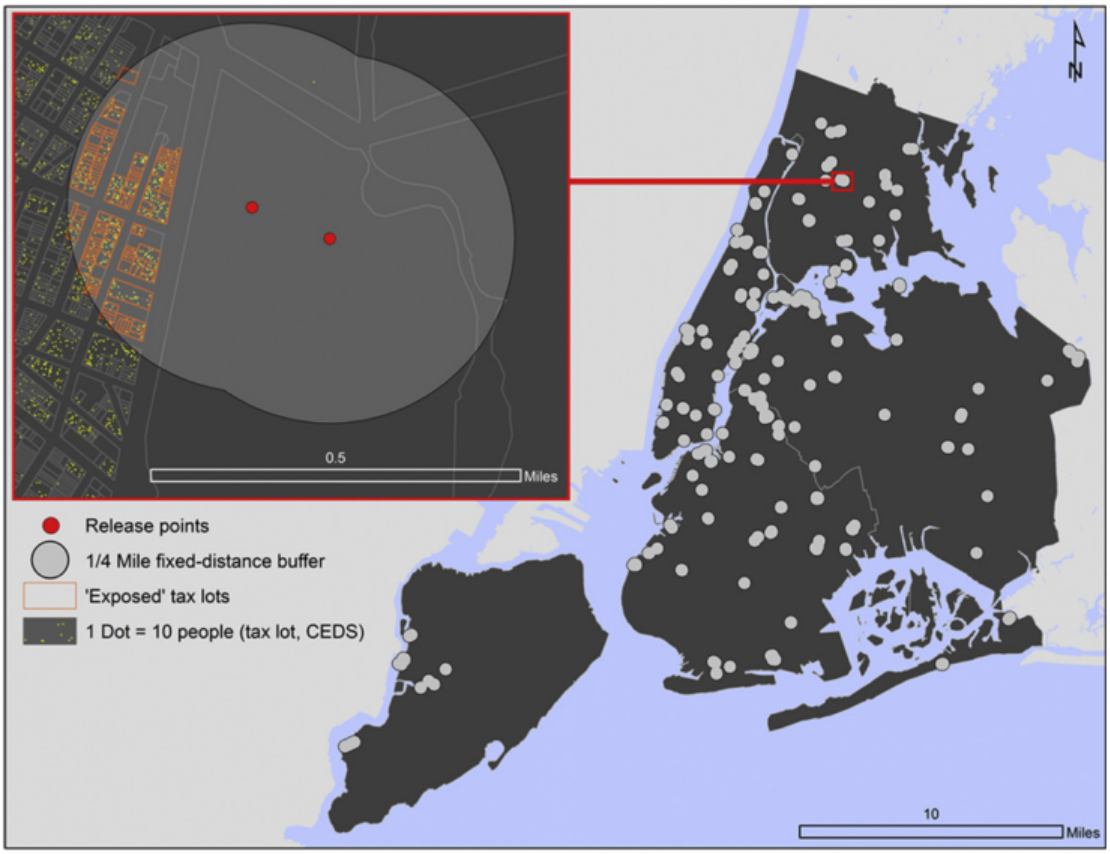


Figure 2.12: Proximity analysis to $PM_{2.5}$ point sources

Figure 2.13 then shows the same area, but now using dispersion modelling.

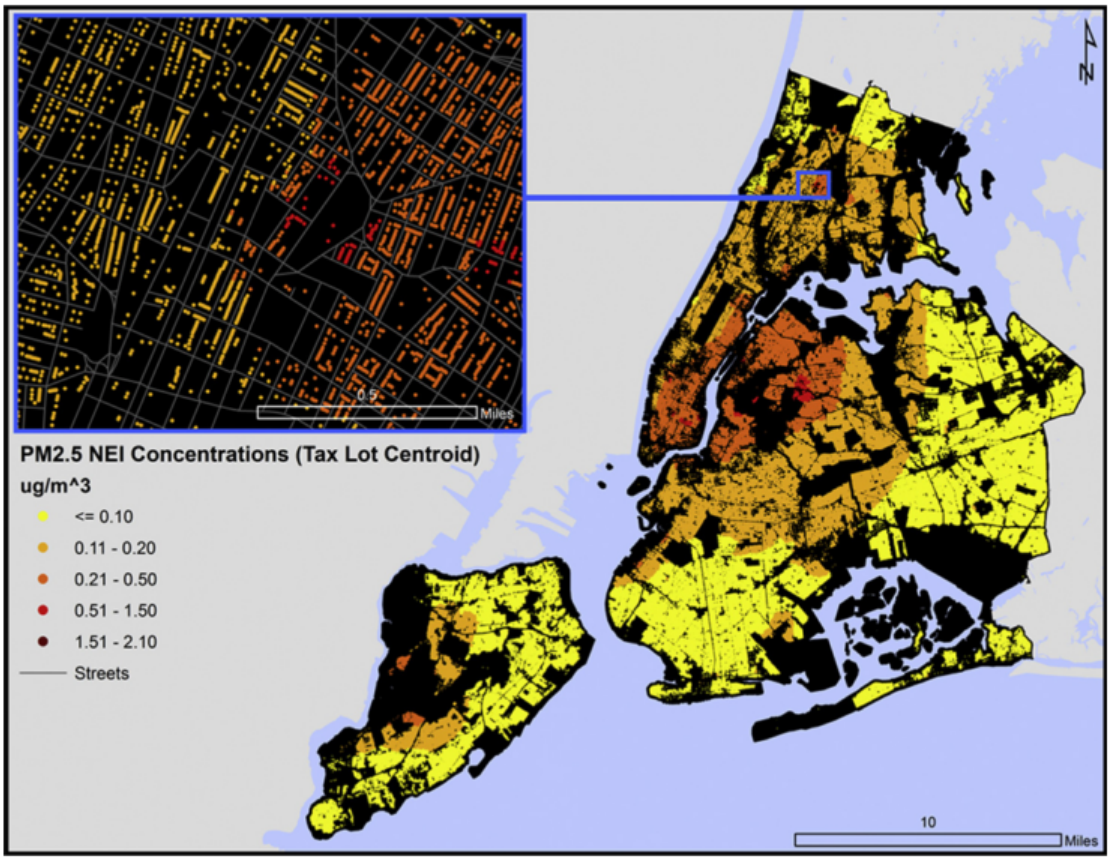


Figure 2.13: $PM_{2.5}$ estimates from dispersion modelling allocated to tax–lots

As is clear to see, the dispersion modelling approach provides a greater degree of spatial detail and clarity. Using this second approach to exposure assessment, they were able to identify that Latino groups in the Bronx and Brooklyn were being dis-proportionally exposed to higher levels of poor air quality than other ethnic groups, which was not evident from the proximity analysis approach. Also using a dispersion model but in London, Tonne et al. (2010) calculated the London annual average concentrations for 2001 and 2005 (pre and post the Congestion Charging Scheme (`http://www.tfl.gov.uk/modes/driving/congestion-charge`)) for NO_2 and PM_{10} on a 20m x 20m grid, and then using this aggregated the data to Ward level (shown in Figure 2.14).

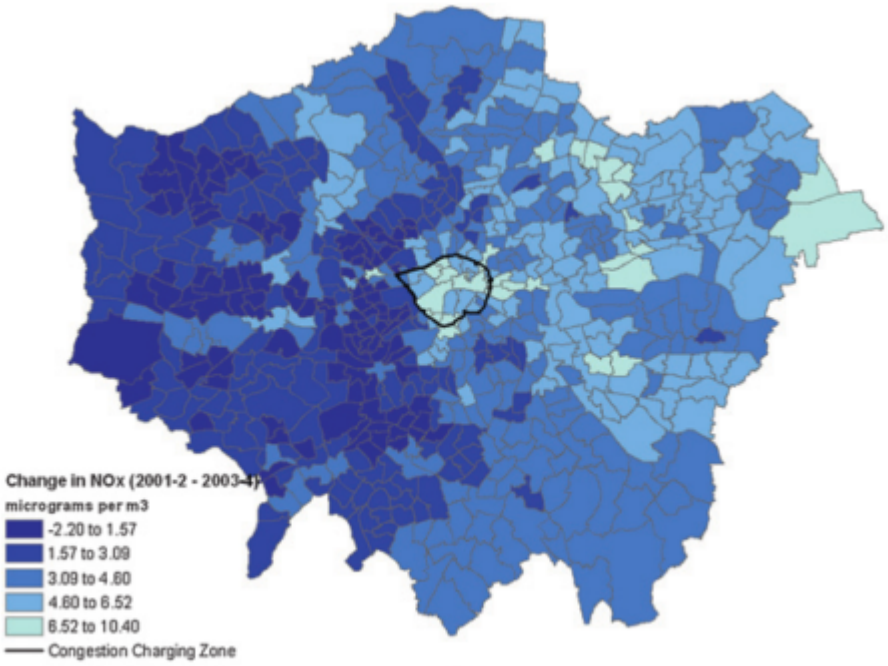


Figure 2.14: Map of NO_x change at Ward level between 2001 and 2005, based on dispersion modelling

The association between this change in concentration and respiratory hospital admissions was then calculated, although no conclusive relation was found.

Both of these studies (Maroko (2012) and Tonne et al. (2010)) seem to be a step in the right direction for improving estimates of exposure as they provide air quality data at a higher resolution than many of the studies that were considered earlier, however there are still concerns that they misclassify exposure by not adequately taking account of temporal variation in air quality, where the subjects are actually spending their time, and the micro-environmental aspects.

2.3.5 Land-use regression

Land-use regression (LUR) maps of air quality, and linking the concentrations from these maps to subjects for health studies, was first undertaken as part of the 'Small Area Variations In Air quality and Health' (SAVIAH) study by Briggs et al. (1997). Land-use regression models combine monitoring of air pollution at a small number of locations, and then develop models using predictor variables normally obtained through geographic data i.e. proximity of roads, land-use, number of nearby buildings, heights of buildings etc. These models are then applied to un-sampled locations in the study area, and concentration values generated using the characteristics of the new location.

A review of the use of LUR modelling for outdoor air concentration values was published by Hoek et al. (2008). They found that the models could be applied relatively successfully to model annual mean concentrations in various geographical locations, for various pollutants including NO_2, NO_x, $PM_{2.5}$ and VOCs, and that the method was better than other geo-statistical methods such as kriging and dispersion methods. However the models were not as effective when a finer temporal scale was required or desired. Dons et al. (2013) in a study accross Flanders (Belgium) used aetholometers with a 5-minute resolution to measure black carbon at 63 locations continuously for seven days. When they compared these values to a LUR model they concluded similarly to Hoek, that existing LUR models were not adequately representing the exposure of the local population, due to the lack of temporal variability in the model. Dons followed this through by attempting to develop an hourly LUR model; with some success. The $R^2$of their models varied between 0.15 and 0.79 according to the time of day and the variables used as input.

In the recent paper by Pedersen et al. (2013), which is an output from the ESCAPE (European Study of Cohorts for Air Pollution Effects) project, a LUR model was generated for 12 European countries, and then temporally adjusted using hourly profiles from nearby monitoring sites (to try to introduce temporal variability that has been lacking from this approach as noted by Hoek). Links between the concentrations at the maternal address of 74,178 women, who had singleton deliveries between 1994 and 2011, were then examined. They found that a 5 μg m^{-3} increase in concentrations of $PM_{2.5}$ during pregnancy was associated with an increased risk of low birthweight.

To summarise, LUR models are being used for health studies, and conclusions between health outcomes and air pollution are being drawn, however there are similar problems with the exposure estimates as in many of the other previous static exposure estimate methods, including that there is no estimation of the amount of time that people spend away from their home and the pollutants and concentrations that they are exposed too i.e. how much time did the mothers in the ESCAPE study actually spent at their maternal address? Studies using LUR models also don't tend to conduct any micro-environmental modelling of the time that subjects spend indoors or in transport, and the methods behind the temporal variability might be considered further i.e. is using a nearby monitoring stations daily variation sufficient to reflect the daily variation at the address point. A further more generic problem with using LUR models for exposure assessment is that they require monitoring to have been conducted in the area that the model is proposed to be used in (and setting up a dense enough network of monitors to provide accurate model results is expensive and time–consuming).

2.3.6 Progressing on from static exposure studies

In the review article *"Spatio-temporal epidemiology: principles and opportunities"*, Meliker and Sloan (2011) discuss how estimating exposure is a rapidly evolving field, and how geographic information sciences, computing power and big data have started to overcome the issues that traditional spatio-temporal epidemiology has often struggled with. As air quality modelling efforts, and linked static exposure assessments, are producing ever-more spatially and temporally accurate estimates of pollutants, spatial analysis techniques and big data are stepping in to compliment these fields by incorporating the mobility of the population in ways not previously possible.

> "We expect exposure assessments to increasingly incorporate space–time dynamics in particular mobility and environmental contaminants, such that it becomes commonplace in the near future" Meliker and Sloan (2011)

The following section titled 'Dynamic exposure and health studies', looks at these new type of studies that explicitly seek to take account of the movement of individuals through different environments, and their exposure in those environments.

2.4 Dynamic exposure & health studies

As we have seen from the studies reviewed thus far, effectively and accurately quantifying exposure to air pollution is fraught with problems. Exposure methods that include sufficiently large numbers of subjects often find it difficult to accurately assign exposure; as the subjects spend time in many different environments, travel between these micro-environments, and often the air quality data is of insufficient temporal or spatial quality. Then the epidemiological studies that are based upon these exposure estimations may therefore be arriving at incorrect health associations and conclusions. Brauer et al. (2002) measured personal exposure and compared it to monitoring site exposure for $PM_{2.5}$ for 16 subjects in Canada and found that in 13 of the 16 subjects the measured ambient concentrations were underrepresenting exposure. Ashmore's review of literature focusing on children's exposure found that their exposure is closely related to concentrations in the home, at school, and in transport, differing significantly from adults concentrations and general outdoor concentrations such as those at monitoring sites (Ashmore and Dimitroulopoulou (2009)).

Although the links between health and air pollution are fairly clear, as discussed in section 2.2 (Health effects of air pollution), it might be that this miss-classification error is biasing any calculated risk factors, regressions or coefficients towards the null value (no association) i.e. once exposure classification is improved the relative risks of increases in exposure to pollutants may be greater than previously thought (Armstrong (1990)).

2.4.1 Personal Monitoring

To gain further insight and a more accurate estimate of individual-level exposure to pollutants, personal monitoring methods can be used. Studies that are described as personal monitoring vary in their use of equipment, and measuring of different pollutants, but generally describe studies that attach portable pollutant monitoring devices to a person or persons, who then go about their normal lives while the devices collect data. The devices normally collect data at short time intervals i.e. one minute, and are combined with location devices like a Global Positioning System (GPS). By using these devices in combination researchers are able to collect data on concentrations at the the places that the subjects are in, and see how levels vary between them. Using such equipment individual level objective direct measures of exposure can therefore be collected. These studies are often considered the "gold standard" of exposure assessment (Ashworth et al. (2013) , de Nazelle et al. (2008)).

Steinle et al. (2013) conducted a review of personal exposure studies of this type. They discussed the difference between traditional exposure assessments using fixed air quality network sites, single micro-environments, and static populations - compared to the new developments in sensor technology that have enabled researchers to directly monitor pollutants while people move through varying concentration fields and activity space. They summarised the personal exposure approach with Figure 2.15 which visualises the joining of location data and air quality data

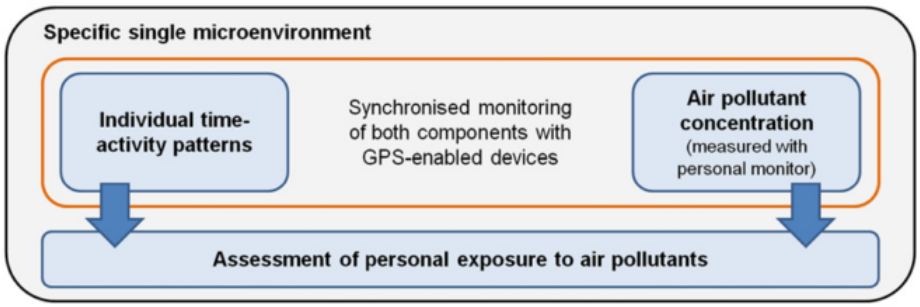


Figure 2.15: Conceptual model illustrating the traditional approach for the assessment of personal exposure to air pollution

This area of air pollution exposure research has, as Steinle notes, accelerated in popularity recently due to the advancements in related technology. The devices have become cheaper, lighter and easier to collect/process data from. The ubiquitous nature of smart phones has also contributed in terms of it becoming the norm for people to carry about technology and have elements of their lives tracked. Indeed, studies such as de Nazelle et al. (2013) have used data from smartphone applications that are not designed for exposure assessment, but which offer a rich data set e.g. the application CalFit for estimating peoples movement style i.e. running, cycling, walking. A search of the terms ('personal exposure' AND 'air pollution') in PubMed, summarised by year of publication in Figure 2.16, illustrates the increase in these studies (whilst acknowledging that there are mitigating circumstances such as the popularity of the field as a whole, and numbers of journals in this area etc.):

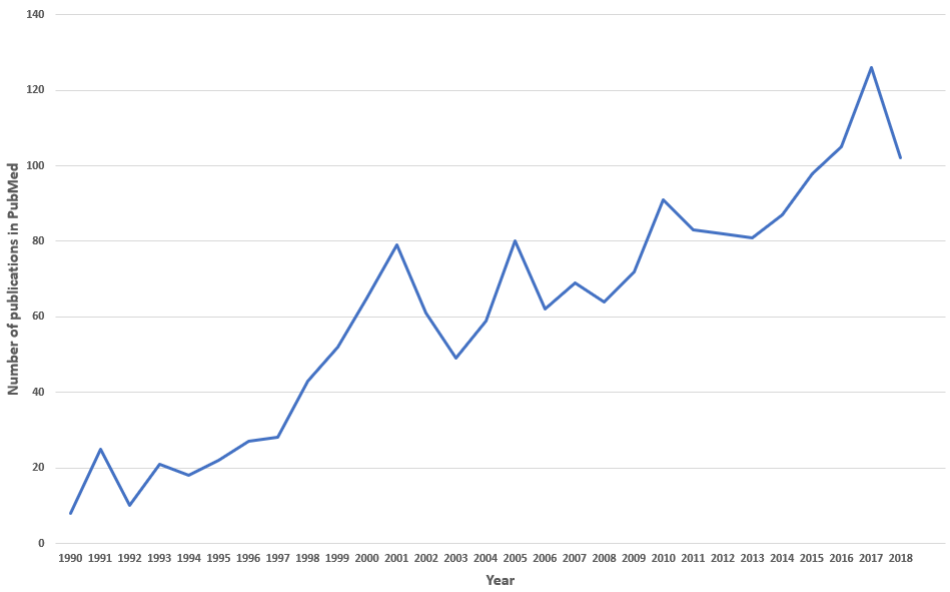


Figure 2.16: Numbers of publications about personal exposure to air pollution in PubMed

Dons et al. (2011) in Belgium completed personal monitoring campaigns for 16 couples. One member of each couple was identified as a full-time worker, and one as a homemaker. Each couple's personal exposure was measured over 24 hours by them carrying a micro-aetholometer and a PDA device which contained a GPS chip for location and time-activity diary for contextual information (which was subsequently completed by the individuals). Very different patterns of exposure were observed between the couples, despite them living in the same place i.e. two individuals living in the same location would have the same exposure according to some of the approaches in the static studies section. Exposure differed by upto 30% between couples, with exposure during transport being identified by Dons et al. as the most important factor in this discrepancy. In a similar study in Italy, Buonanno et al. (2014) conducted personal monitoring on 24 non-smoking couples and measured their exposure to ultra-fine particles using a Phillips NanoTracer (which included a GPS). In this study the couples were all male-female, the male being a full-time worker and the female being a homemaker. Given the emphasis that Dons found on transport, it was slightly surprising that this study found women to have higher exposure than men. This was attributed to the emissions from cooking stoves in the home and that the particles stay in the environment for prolonged periods. Although the two studies actually consider different metrics (black carbon compared to ultrafine particles) so the comparison is hard to make directly – there are not many sources of black carbon in the home, but transport is a major source outside of the home. In a further study Broich et al. (2011) used a GPS, and GRIMM 1.09 monitor to measure particulate matter (in various sizes) for sixteen people over 24 hours each in Italy.

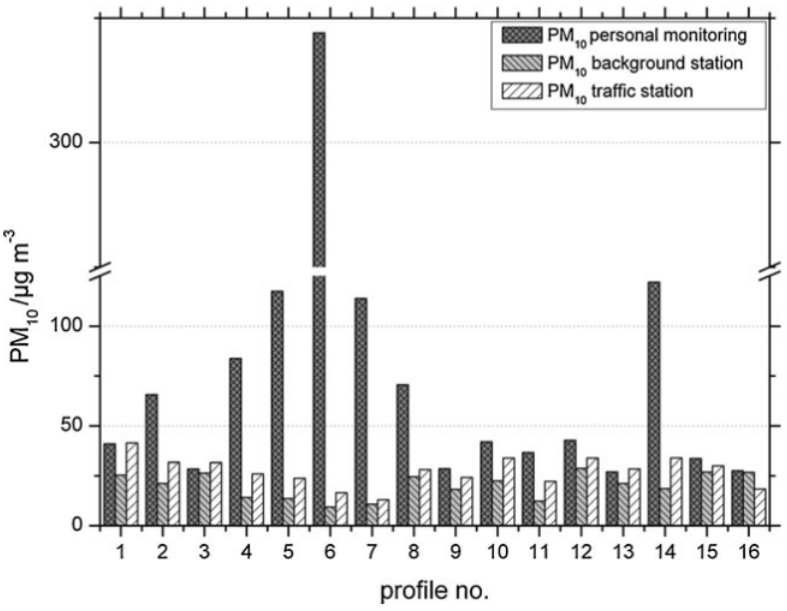


Figure 2.17: Average exposure over 24 hours to PM_{10}, comparing personal monitoring, background monitoring stations and traffic monitoring stations

As can be seen in Figure 2.17 from Broich et al. (2011), there were large variations between the personal exposure concentrations, and concentrations at nearby monitoring sites. The averages over 24 hour between participants varied from 27 to 322 μg m^{-3}, and like Dons they found that exposure levels were heavily dependant on the travel participants undertook, and the micro-environments they spent time in. Broich argues this is why direct personal exposure monitoring is needed for accurate exposure assessments rather than proxy methods (monitoring sites or address). Although for epidemiological studies that might relate patterns from monitoring sites to patterns in incidents of poor health, this may not necessarily the case. If the personal exposure results for instance are always say 40% than the monitoring site, then the monitoring sites would still capture the trends and might be able to relate these trends to hospital admissions.

While these types of study are excellent at providing hyper-local and time resolved personal exposure data, they tend to struggle to provide data that is useful for considering alongside the harmful health effects of poor air quality. The studies give results for individuals or small groups of people which are not able to be used alongside health-related epidemiological data such as hospital admission records or incidences of asthma in a certain region. As Chaix et al. (2013) argues, improving measuring of exposure to environmental conditions by accounting for the movement of individuals is critical, however by using such small numbers there is a danger of over-generalisation between exposure and health. Measuring 20 school-children's exposure to PM_{10} for example, and then making wide-ranging assumptions about the health effects of PM_{10} on all school children seems overly simplistic. It might be that those 20

children are in the top 5% of exposures and not a good representation of school children in general.

Steinle et al. (2013) argues that the trend in the area of calculating accurate exposure estimates for large populations is for the development of the personal monitoring approach, i.e. distribution of low-cost, accurate GPS devices combined with accurate low-cost unobtrusive portable air quality measurement devices. In this way the amount of data collected through this 'gold-standard' method would be adequate to allow extrapolation to large groups of people i.e. if 80% of school children carried a monitor for 12 months, then concerns about how representative the data is would be less valid. A model of their theoretical approach is shown in 2.18.

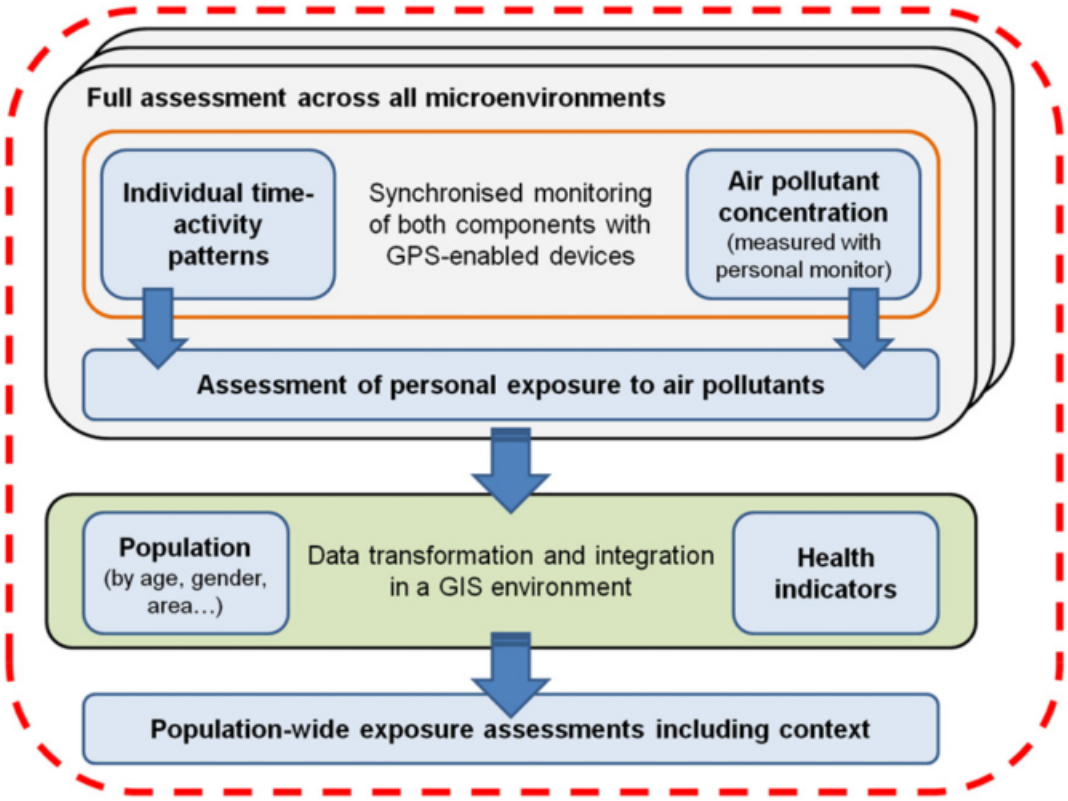


Figure 2.18: Conceptual model for the assessment of individual and population-wide exposure to air pollution including effects

Minguillón et al. (2012) seemingly agree with this large-scale personal monitoring approach, or at least with the notion that personal monitoring is required to properly understand exposure. In their study on pregnant women's exposure in Barcelona in 2009 they compared measurements from the subjects household balcony, inside their house, and the data from personal monitoring equipment. They found mean concentrations for $PM_{2.5}$ of 20 μg m^{-3}, 24 μg m^{-3}, and 27 μg m^{-3}, for outdoor, indoor, and personal samples respectively. They

concluded that it was important to rely on personal exposure measurements for epidemiological studies.

Against this approach however is the practical nature of conducting these monitoring campaigns. The technology for this type of study is not yet available (as the above authors note). In addition, even if it were, the distribution of the devices and automated collection of the data on a large enough scale would seem difficult to achieve. Though this may change with the advent of smaller and lower-cost measurement technology. What personal monitoring studies do show researchers in this field is where subjects the majority of their time (indoors), and where they are exposed to the highest levels of air pollution (normally during travel). By doing so they emphasise how important these factors are in understanding the true exposure of individuals compared to current or traditional methods.

Developing a modelling approach, with a focus on the accuracy of the modelling in these environments to improve exposure estimates, seems a more practical approach to better understanding population exposure.

2.4.2 Infiltration

Infiltration refers to the diffusion of outdoor air into the environment inside a building. The amount of outdoor air that ingresses depends on ventilation, air conditioning and on the indoor–outdoor temperature gradient. EU guidelines now state that buildings must be insulated to save energy, however as buildings allow less air to be exchanged with the outside environment, the indoor concentration of pollutants can increase if there are significant indoor sources. Thus increasing the air-tightness of buildings can have negative impacts on health (Gens et al. (2014)). This is offset of course by meaning that harmful outdoor-generated pollutants will not as easily be able to infiltrate into the indoor environment. The study of indoor air has steadily become an inherent part of modern exposure research (Steinle et al. (2013)). Figure 2.19 below illustrates how outdoor air infiltrates buildings and mixes with indoor air (from Chen and Zhao (2011)).

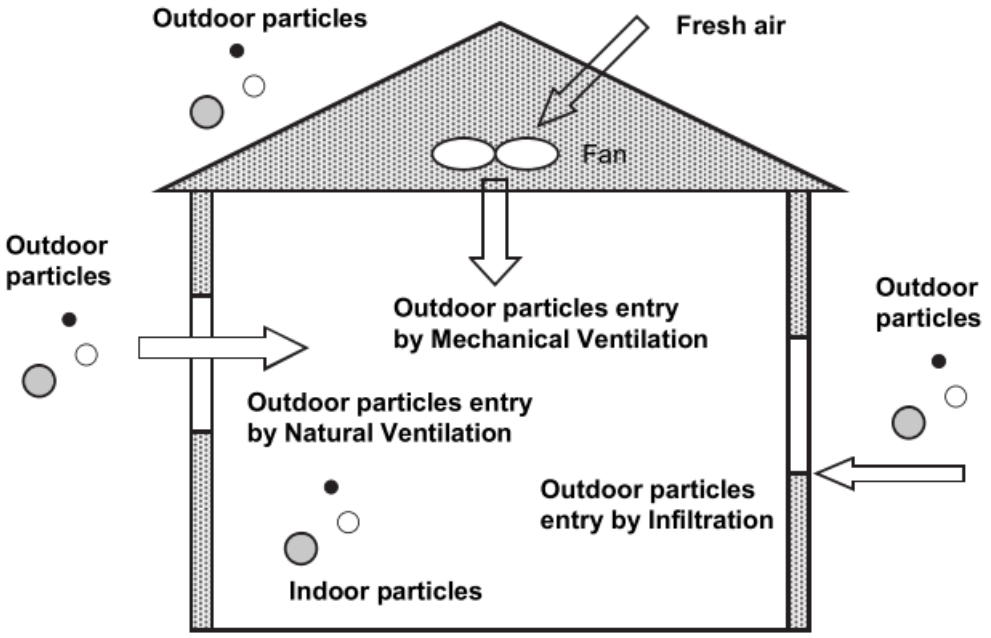


Figure 2.19: Infiltration

Indoor pollutants have many sources such as cigarette smoke, cooking, heating and cleaning products. They can substantially change a persons exposure (compared to ambient concentrations), and characterisation of them is difficult. To consider the exposure of subjects to air pollution over days/weeks/months, indoor air is an important consideration. A better understanding of the factors influencing infiltration and indoor sources can improve exposure assessment methods and contribute to reduced exposure miss-classification in epidemiological studies (Colbeck and Nasir (2010), MacNeill et al. (2012)). Doing so depends on the metrics that are being used for considering the health effects e.g. to better understand the effects of traffic pollution on health, indoor sources (cooking, heating) would not be needed and the infiltration of outdoor air into the indoor environment would be key. But for a model which considered exposure to all pollutants then all sources would be needed.

To try and better understand the differences between indoor and outdoor air quality, as part of a large cohort exposure study in Ontario (Canada), MacNeill et al. (2012) measured concentrations of $PM_{2.5}$, ultra-fine particles (UFPs), black carbon and humidity both inside properties and in the back garden of properties simultaneously for two weeks. As is common with this type of study the results were discussed in terms of indoor/outdoor ratios, otherwise known as I/O ratios. This is the concentrations inside the property, divided by the concentrations outside of the property. So if a property had a $PM_{2.5}$ I/O ratio of 0.5 at 10am and the concentrations outside were 10 μg m^{-3}, then this would mean that concentrations inside the property were 5 μg m^{-3} (10 multiplied by 0.5).

The median daily estimates found in the study ranged from 0.26 to 0.36 across seasons for $PM_{2.5}$, with the ranges typically related to window-opening behaviours, air conditioning, meteorological variables, home age, use of electrostatic precipitators and stand-alone air cleaners. The determinants of indoor source concentrations were related to cooking, candle use, supplemental heating, cleaning, and s of people in the home. Expanding on the last of the variables noted by MacNeill, Colbeck and Nasir (2010) comments on the close relationship between indoor concentrations and human activities, stating that humans are responsible for their own 'personal cloud' i.e. exposure to airborne particles resulting from their own activity.

Challoner and Gill (2014) looked at $PM_{2.5}$ and NO_2 concentrations in ten different city centre buildings in Dublin. They found $PM_{2.5}$ I/O ratios close to 1 (similar to outside air) for the ten commercial buildings, however after studying the temporal variation in these levels they suggested that indoor sources and/or re-suspension of $PM_{2.5}$ seemed to have a more significant impact compared to variations in outside air quality. Kearney et al. (2014) measuredPM continuously for seven consecutive days in 74 Edmonton (Canada) homes in 2010. Simultaneous measurements of outdoor (near-home) and ambient (at a central site) concentrations were also measured. As with the studies above, they found considerable variability ranging from 0.10 to 0.92 in winter and from 0.31 to 0.99 in summer.

Given the number of variables that seem to contribute to variations in indoor air quality, making broad-sweeping assumptions to create inputs to epidemiological models seems difficult. However WHO calculated in 2005 that people spend 89% of their time indoors, Lai et al. (2004) found a figure of (89.5%) from their research, and Schweizer et al. (2007) found similar - 20.66 hours (86%). Efforts to better quantify this exposure relationship and create as accurate a picture of exposure are therefore needed. The variability in I/O ratios within and between homes may cause substantial exposure miss-classification compared to only using ambient measurements (Kearney et al. (2014)).

A dynamic exposure model i.e. one that is going to take account of different micro-environments where people spend their time, needs to attempt to model concentrations indoors. People spend such a large percentage of their time indoors that this environment needs 'adjustments' from ambient concentrations. The following paragraphs therefore consider a few recent studies that have attempted this.

MESA-Air stands for the "Multi-Ethnic Study of Atherosclerosis and Air Pollution" and was a large research project based in Washington State (USA). The project was designed to examine the relationship between air pollution exposures and the progression of cardiovascular disease (Allen et al. (2012)). To do this the researchers needed to quantify exposure to indoor air, and so they aimed to develop models to predict I/O ratios for around 6,000

homes. They did this by collecting 526 two-week, paired indoor-outdoor $PM_{2.5}$ filter samples from a subset of homes in their study. Taking account of specific weather and seasonal variables, as well as using information from questionnaires, they made a regression model which predicts I/O ratios (mean of 0.62, SD of 0.21) using easily obtained variables. This is an important study in the cross–over area of air pollution exposure assessment, indoor–outdoor concentrations and epidemiology as it was the first study with a large number of subjects to incorporate variation in residential exposure into exposure assessment. The effect that this new information made to epidemiological study of the prospective cohort is not published at the time of writing.

Hussein et al. (2014) took a similar approach with their exposure model, however they also considered dose (Figure 2.20). Their model calculates exposure by considering infiltration into the indoor environment, indoor sources, building types, building sizes and the time-activity pattern of the individual. A mass-balance model is used for the indoor exposure which not only models infiltration but the indoor sources.

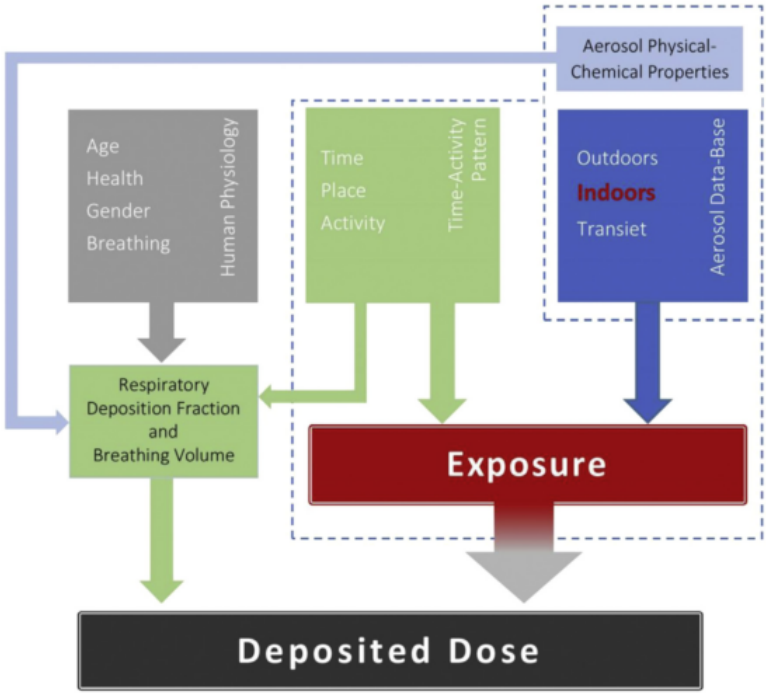


Figure 2.20: Exposure model incorporating indoor exposure

They illustrate the use of the model by calculating exposure and dose for 24 hours for a test individual, showing the inputs needed for the various model parameters. The advantage of this approach is that it starts to show a way for indoor exposure to be properly included in an exposure assessment model, however as the authors acknowledge the detail and amount of data required for input to the model (i.e. volumes of buildings, detail of indoor sources,

time activity patterns) is not yet available on a large enough scale and robust enough, to expand this approach to large numbers of subjects. Lacking exposure modelling of the journeys between the environments would also seem to be necessary for this model to give a more holistic picture of exposure. However the inclusion of the calculation of dose in such a dynamic exposure model is clearly a positive step. Dose being a quantification of the amount of pollutant(s) that the subject actually breathes in and makes its way to the airways and lungs of the subject as per the discussion and diagram of Brook et al. (2010) in Section 2.2.1.

Mölter et al. (2012) also developed a dynamic exposure model that considered pollutant levels in indoor micro-environments, specifically focusing on NO_2. They had 60 school-children (ages 12–13) wear personal monitors for two days and while doing so keep a detailed log of their activity, location, and the activity of others around them that might generate pollutants e.g. their parents cooking or smoking at home. Using the time-activity patterns, modelled outdoor concentrations and I/O ratios from literature (e.g. 2 for cars/buses, 0.5 for school) they calculated exposure for the set of children and compared it to their monitored data. The results of this micro-environmental exposure model agreed well with the personal exposure measured by the children, however a great deal of contextual information was required to come to such close agreement and it is hard to see how a model of this type could be applied to a much larger cohort of subjects without the same level of detail available. Given this the transfer-ability of the study to other subjects and areas is not useful as a tool in and of itself, but it does show that good agreement can be obtained between modelling and monitoring of personal exposure given sufficient data. An additional problem with using this approach elsewhere is that the air quality data was based on a land-use regression model, which as was discussed in Section 2.3.5 (Land-use regression), requires extensive monitoring to be accurate and often lacks temporal variation.

As discussed earlier in this research, the health effects of exposure to air pollution are not yet fully understood. Studies have tended to assign exposure based on outdoor concentrations at the subjects residence using long term average concentrations or on monitoring stations concentrations for studies of short-term exposure. Given this it is no surprise that the negative effects of poor indoor air quality on a populations health are similarly not yet understood (although the approaches discussed in the preceding paragraphs are contributing to a better of understanding of overall exposure i.e. they are starting to contribute to understanding and quantifying exposure indoors, while other studies that quantify exposure in other environments are completing the picture).

Gens et al. (2014) is one of very few studies that have attempted to consider how indoor air is affecting health, specifically looking at the increase in air-tightness of modern EU build-

ings. The assessment was based on modelling exposure to fine particles originating from both outdoor and indoor air, including environmental tobacco smoke. Exposure response relationships were derived and the results showed an increase of adverse health effects in all considered countries (ranging for health effects from 0.4% in Czech Republic to 11.8% in Greece for 100% insulated buildings) due to an accumulation of particles indoors. Unsurprisingly considering only the effects of outdoor air led to a decrease of adverse health effects. Although the conclusions drawn for this response relationship seem in doubt as the odds-ratios have been applied to the combination of indoor and outdoor air, when they are only designed and suitable to be used for outdoor air.

Chen and Zhao (2011) conducted a review of modelling approaches on the relationship between indoor and outdoor particles and summarised that I/O ratios vary considerably due to the difference in size-dependent indoor particle emission rates, the geometry of the cracks in building envelopes, and the air exchange rates. Concluding that due to this it is difficult to draw uniform conclusions on I/O ratios, as they vary so much between building types.

The studies discussed in this section show that indoor air varies considerably from outdoor air, and that research is ongoing as to how to effectively model this as part of a holistic exposure model that also incorporates other environments such as different transport modes. Research so far has identified that indoor air quality is effected by outdoor concentrations, the number and activity of people inside the building (particularly apparent for particulates), the ventilation systems (or lack of), indoor sources, meteorology and the air-tightness of the building.

2.4.3 Transport

As briefly discussed in Section 2.1.6.2 ('In-vehicles'), air quality when in transport can differ greatly from ambient concentrations. Table 2.3 (also from Section 2.1.6.2) showed measurements in different transport modes which were different (mostly higher) than that of the ambient concentrations. A review of pollutant concentrations in different transport modes by Karanasiou et al. (2014) provides an excellent resource for considering the breadth of the differences between in-transport air and ambient air. Reflecting transport exposure as part of a persons total exposure is important as it is thought that individuals gain a significant contribution of their daily exposure during travel, despite the small percentage of time that is typically spent doing so. For black carbon exposure Dons et al. (2011) found subjects spent 6-8% of their day in transport, where they accumulated 21% of their exposure. Failing to account for these important exposure events contributes to exposure miss-classification.

As can be seen from the data compiled by the Office for National Statistics (Office for National Statistics (2010)) in Figure 2.21, the three most popular modes of transport in England and Wales over the last half a century have been car and van, followed by bus and coach, and finally rail.

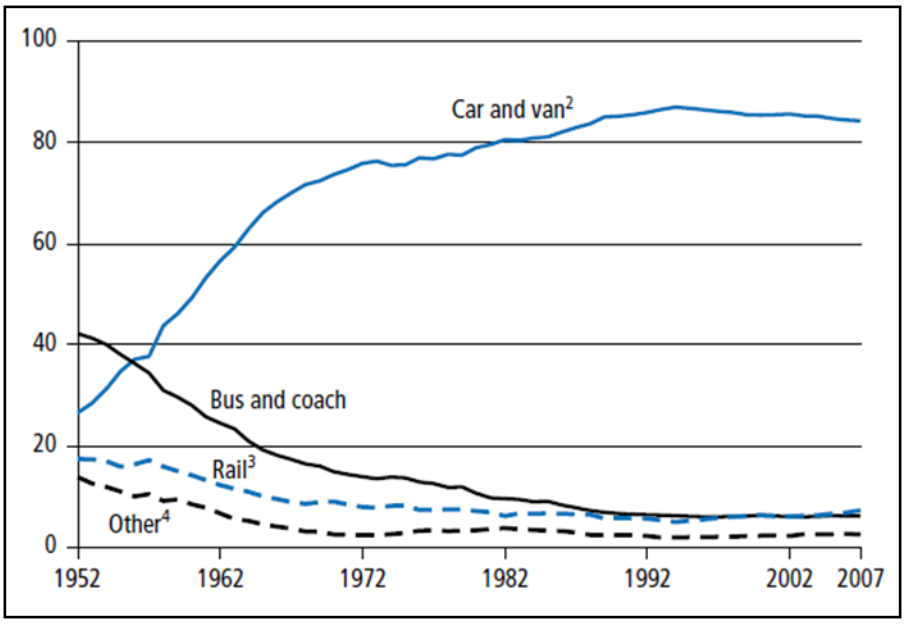


Figure 2.21: Transport profiling from Office for National Statistics (ONS) census data between 1952 and 2007

Studies which have sought to measure and understand exposure in these transport modes are now discussed, with the addition of cycling. Cyclists are included as a special case due to their increased inhalation rates and proximity to road traffic which may significantly increases their exposure and dose of poor air quality.

2.4.3.1 Bus and Coach travel

The review by Karanasiou et al. (2014) found that pollutant concentrations inside buses and coaches vary greatly by country, likely reflecting the variation in weather conditions, vehicle types, ambient concentrations, and fuel sources of the vehicle/surrounding fleet. In one study concentrations in the Netherlands were found to be higher in diesel vehicles than on the same routes with electric fuelled buses, suggesting that the own vehicles exhaust impacts on the exposure inside the vehicle (vehicle 'self pollution'). Though this could also be reflective of the permeability of the vehicle or the number of times the doors were opened or closed, allowing ambient concentrations to ingress. When compared to ambient concentrations, a Paris study found an I/O ratio of 1.3, and in the Netherlands a similar ratio

of 1.43. Typical $PM_{2.5}$ concentrations inside the vehicle across the eleven studies reviewed were in the range 35 – 69 μg m^{-3}.

A study not covered in the review is by Song et al. (2009) in York, U.K. Measurements were made simultaneously by placing an optical particle monitor in the middle of a bus in York over a period of three days in May 2007, and comparing the data recorded to measurements from an identical monitor on the bonnet of a car which followed the bus (Song et al. (2009)). Measurements were completed during the morning and evening rush hours over 24 hours. Figure 2.22 shows how the relationship between out-bus and in-bus concentrations was positively linear, with a higher gradient as particle size increases (the column labelled 'Dependent variable' is the size fraction of PM). Also having the windows closed on the bus actually led to increased concentrations compared with having the windows open, suggested to be due to re-suspension by passenger activity.

Dependent variable	Mean concentrations		*I/O* ratios		*F*
	In-bus	Out-bus	Window open	Window closed	
0.75–1.0	2200	1129	1.53	2.16	27.4
1.0–2.0	1514	705	1.55	2.48	25.4
2.0–3.5	819	301	1.78	3.30	23.9
3.5–5.0	332	45	4.51	8.95	30.3
5.0–7.5	128	12	6.35	12.51	35.8
7.5–10	44	3	9.89	18.80	59.6
10–15	28	1	24.13	34.49	95.0

Figure 2.22: Comparison of mean in-bus and out-bus particle concentrations

I/O ratios for PM in the size range of 0.75 to 2.0 are 1.53 and 1.55 for when the windows are open, and 2.16 and 2.48 for when the windows are closed. There are many other factors which complicate drawing conclusions between studies, but it is interesting to note how closely the former of these ratios agree with the studies in Paris and the Netherlands for similar PM fractions discussed at the start of this section.

Using data from this study, the Song et al. (2009) paper concludes by compiling a model for estimating the indoor/outdoor ratio of different PM fractions. The model showed broad agreement with measured particle concentrations inside buses and demonstrated that re-suspension by passenger activities and deposition to the surface of the passengers had significant effects on the concentrations.

The health implications of bus and coach exposure is not discussed by the Karanasiou review or the Song paper as a separate issue or metric. Being able to include exposure by this transport mode in a large-scale exposure assessment exercise is required for exposure mis-classification to be reduced. Though is further complicated, as in many of the sections

of this report, by the composition of particles i.e. particles from different sources have more or less harmful effects on health.

2.4.3.2 Car travel

Studies on exposure while travelling inside cars are more frequent than those of other transport modes. From reviewing studies of car exposure Karanasiou et al. (2014) found that typical European particulate matter concentrations inside passenger cars were in the range of 36-76 μg m^{-3} for PM_{10} and 22-85 μg m^{-3} for $PM_{2.5}$. Whether the car itself was powered by diesel or gasoline was found in most studies to make little difference to particulate matter, particle number counts and black carbon. However Jalava et al. (2012) found that pollutant concentrations inside the car depended highly on the type of fuel used.

A separate review by Nasir and Colbeck (2009) found that exposure to particulate matter in the car micro-environment is largely dependent on traffic congestion, road network layout, vehicle design/condition and ambient concentrations.

Dons et al. (2013) studied slightly different parameters, or ways of considering the parameters, of black carbon exposure in vehicles in Figure 2.23. It was noted that in urban areas concentrations inside vehicles are higher compared to exposure in more rural areas; with the same holding true for highways versus local roads for motorists – suggesting that the surrounding fleet and traffic speeds affect in-vehicle concentrations.

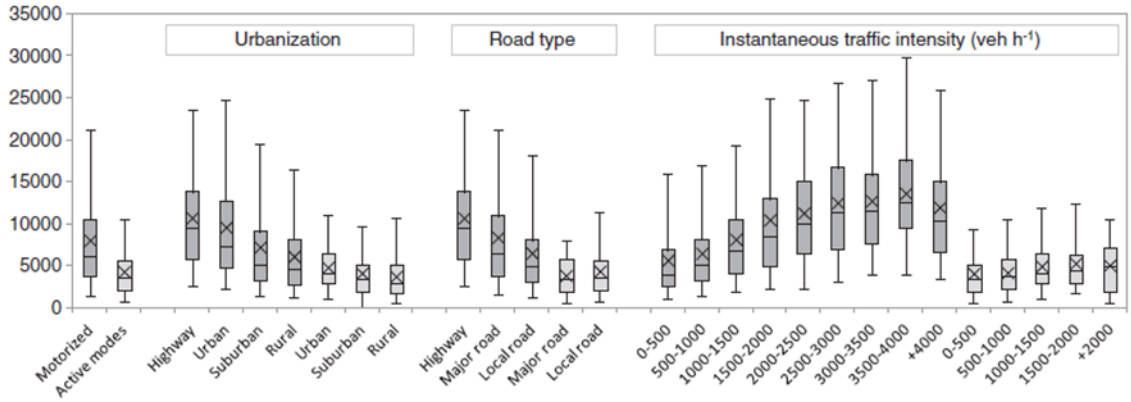


Figure 2.23: External factors affecting in-car black carbon exposure (μg m^{-3}). Vehicles are the dark boxplots. Walking are light

By taking all the factors studied by Dons, a regression model of car I/O ratios for black carbon was developed and is shown in Figure 2.24.

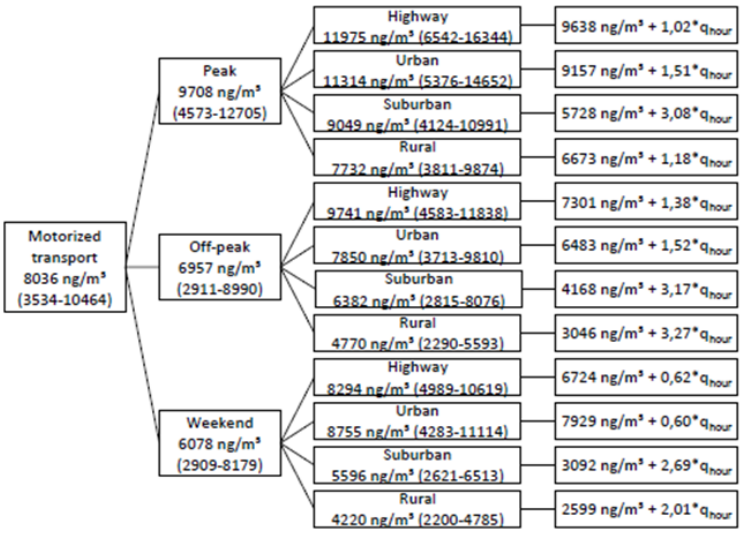


FIGURE SI8: Visualization of the regression model for motorized transport (q_{hour} is hourly traffic intensity). Peak hour is defined as 7-10 a.m. and 4-7 p.m. The tree should be used as a look-up table, e.g. if a car trip in peak hour in a rural area is considered, average exposure of the motorist is 7732 ng/m³. The interquartile range (ng/m³) is in brackets.

Figure 2.24: Transport indoor/outdoor ratios

However the use of these ratios are not well tested and do not agree well with other ratios from the literature. It seems that given other known influences such as ventilation systems that a deterministic model based only on time of day and road type is too simplistic. For example in two studies by Gulliver and Briggs (Gulliver and Briggs (2004) and Gulliver and Briggs (2007)) particulate matter concentrations inside cars in Northampton and Leicester were roughly 30% higher than concentrations at nearby monitoring stations. A table of mean concentrations from the 2004 paper are shown below in Table 2.5).

Table 2.5: In-vehicle PM concentrations

PM Fraction	**Mean (μg m^{-3})**
PM_{10}	43.16
$PM_{2.5}$	15.54

The main conclusions from studies on in-car expose were that exposure is significantly influenced by traffic intensity and ambient air pollutant concentrations (being strongly linked to each other) as well as the choice of ventilation used inside the vehicles themselves. Is is a complicated picture. Estimating the exposure of individuals from the time they spend in the car micro-environment should take into account the factors outlined to try and as accurately as possible reflect reality.

2.4.3.3 Bicycle

Between 2001 and 2011 the number of people living in London that cycled to work more than doubled from 77,000 to 155,000. Over the same time period there were also large increases in other large UK cities; Brighton by 109%, Bristol by 94%, Manchester by 83%, Newcastle by 81% and Sheffield by 80% (Office for National Statistics (2014)). This rise in cycling is normally attributed to a number of factors including local authorities building more cycle-friendly infrastructures, the promotion of cycling as a way to keep fit (supported by the Governments cycle-to-work discount scheme), and the public generally becoming more environmentally-aware.

In some cities bicycle sharing systems have also contributed to this rise in cycling journeys. In London the Barclays Cycle Hire scheme was launched in July 2010 with around 5,000 (now around 11,000) bikes distributed around London at specially installed docking stations (Transport for London (2014a)). Similar systems operate in many places around the World. Although there is no definitive list as new schemes are opening all the time, in 2013 Larsen (2013) estimated there were more than 500 cities in 49 countries hosting advanced bike-sharing programs, with a combined fleet of over 500,000 bicycles (compared to 213 schemes in 2008 (Wikipedia (2014))).

The Karanasiou et al. (2014) review looked at cycling exposure in 20 European studies, mostly in the UK and the Netherlands. They found mean exposure values for $PM_{2.5}$ of 29–72 μg m^{-3} and for PM_{10} of 37-62 μg m^{-3}. The studies based in London (Kaur et al. (2005) and Adams et al. (2001a)) found that cycling exposure to $PM_{2.5}$ depended on the route taken, finding busier routes with more traffic increased cycling exposure. Ragettli et al. (2013) found similar - bicycle travel along main streets between home and work place contributed 21% and 5% to total daily ultra-fine particle exposure in winter and summer, respectively, and that exposure could be reduced by 50% if main roads were avoided. At odds with this, the Netherlands study (Zuurbier et al. (2010)) found no difference between routes. Though in their methods it seems that there was overlap between the low-traffic and high-traffic routes, and also that their low-traffic routes were only low on car traffic and they were frequently used by mopeds, which may explain this difference.

The main variable common in studies of cycling exposure seems to be location. A cyclist riding in the middle of a busy road will be exposed to higher concentrations than on a separate cycle-lane, or a back-street. Given cyclists are not enclosed in any sort of vehicle this direct relationship between air pollutant concentrations and exposure is to be expected, as the cyclist is in closer proximity to the main sources (car exhausts).

With regard to the negative health impacts of exposure to poor air quality while cycling,

unlike other transport modes such as car and bus, a number of recent studies have tried to quantify this. Woodcock et al. (2014) used Barclays Cycle Hire data from Transport for London (TfL) to model the health impact for 578,607 people in London, mostly (78%) aged between 14 and 45. They used the bike hire usage data, modelled the journeys between the docking stations, and combined this with general travel data, data on physical activity levels and road collisions, and finally 20 m by 20 m grids of annual average $PM_{2.5}$ data. Cycling exposure was compared to replacing those journeys with walking or public transport. They found that the population benefits from the cycle hire scheme substantially outweighed the negatives, with a net change of minus 72 daily adjusted life years (DALYs) among men and minus 15 for women (negative DALYs represent a health benefit). In a very similar study in Barcelona Rojas-Rueda et al. (2013) did a Health Impact Assessment (HIA) for eight different scenarios where the the population of Barcelona was presumed to change transport mode from car to either cycling or public transport. Traffic incidents, physical activity and air pollution exposure were then estimated for each scenario and the relative health changes estimated in DALYs. For the scenarios where 20% and 40% of car trips where replaced by cycling trips, there were changes of minus 138 and minus 275 DALYs. Rather than doing a health impact assessment Nwokoro et al. (2012), back in London, did more direct assessments of cycling exposure by looking at the amount of black material in the airways of non-cyclists and cyclists. They found that commuting to work by bicycle in London was associated with increased long-term inhaled dosage of black carbon, however the relationship was difficult to properly quantify. This was because cycling typically takes longer for those people to cycle to work than, for example, a train journey. Route choice further complicates matters as although being away from highly polluted routes lowers exposure, the increased journey time may increase total exposure over the trip. Figure 2.25 (from Nwokoro et al. (2012)) below shows the increased macrophage carbon measured in cyclists compared to non-cyclists (A macrophage is a cell responsible for destroying harmful pathogens).

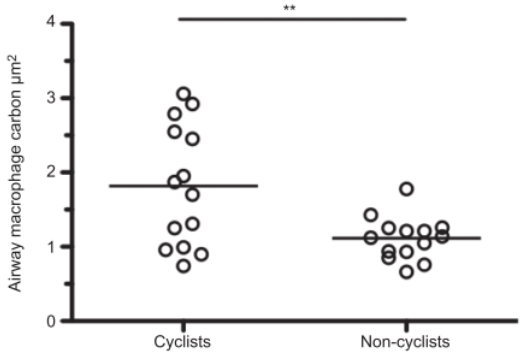


Figure 2.25: Airway macrophage carbon in cyclists and non-cyclists

In summary cycling exposure is very dependant on the routes taken by cyclists, and their position on the road. The negative health effects of cycling have been explored in comparative transport mode studies showing that when taken alongside other variables such as risk of injury, cycling has health benefits. There are a lack of studies that take cycling exposure as part of a larger daily exposure health assessment alongside other transport modes and indoor micro-environments. A number of the health studies are encouraging in that they do model sufficiently large groups of people to draw epidemiological conclusions but only the transport aspect of these people's exposure is explored rather than cycling in the context of their daily typical exposure. Also there are other parts of their models which could be improved, for example the studies in London and Barcelona only used annual average air quality concentrations (rather than daily or hourly for example).

2.4.3.4 Train travel

In their 2014 review of commuter exposure in Europe (which has been referred to extensively in this section on transport exposure), Karanasiou did not review train travel – focusing on cycling, cars and underground subway systems. However Nasir and Colbeck (2009), Colbeck and Nasir (2010), Dons et al. (2011), Ragettli et al. (2013) and Knibbs et al. (2011) all have elements of their publications which take train travel exposure as a separate travel micro-environment.

Nasir and Colbeck (2009) and Colbeck and Nasir (2010) found concentrations during peak journey times in air-conditioned carriages were 44 μg m^{-3}, 14 μg m^{-3} and 12 μg m^{-3} for PM_{10}, $PM_{2.5}$ and PM_1 respectively, but that during off-peak times, concentrations were about half this. Concentrations were lower again in non-air conditioned carriages. They drew the conclusions that particulate levels inside trains are strongly influenced by the numbers of passengers and that their movement and presence causes particle re-suspension. Peak travel times (when there are more passengers) coincided with higher particulate concentrations. Figure 2.26 (from Nasir and Colbeck (2009)) shows the effect of train stops on concentrations in the train cabin. It would have been interesting to see simultaneous outdoor concentrations to see how they varied to the indoor concentrations and more detailed information on passenger numbers. It might be that outdoor concentrations contribute to a certain percentage of the indoor concentrations, which are then varied positively or negatively by passenger numbers.

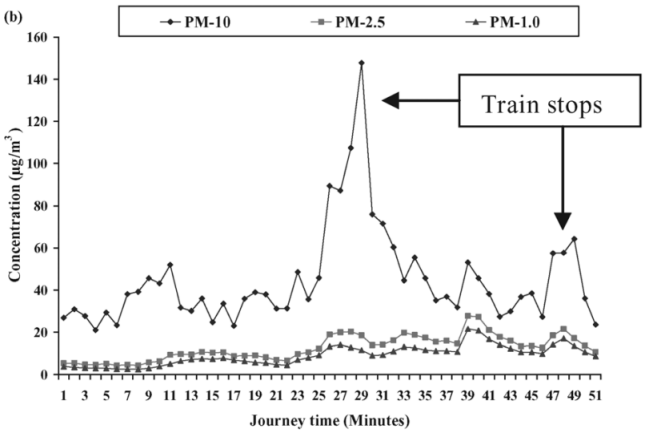


Figure 2.26: Train stops and train exposure

Knibbs et al. (2011) reviewed articles to look at the difference between diesel and electric powered trains, and based on the limited data available found that the power source of the rail vehicle appears to strongly affect ultra-fine particle concentrations (diesel powered trains have higher in-train particle concentrations than electric trains). High concentrations of particles are not just confined to the trains themselves either, Ragettli et al. (2013) found UFPs levels within train stations twice as high as suburban background and nearby residential streets in Basel, Switzerland.

In summary the variables that drive exposure levels on trains are the fuel source, the number of stops that the train makes, the influx of passengers at those stops, and the concentrations outside of trains. No research on specifically how train travel exposure affects the health of passengers was found. As with the other transport modes reviewed thus far, a model which incorporates these variables alongside other micro-environments should be able to develop a more accurate picture of human exposure over days or weeks.

2.4.3.5 Underground subway systems

In many worldwide cities the population travel between locations using metro systems. The London Underground was the first such system to open in 1863, but other cities have followed including New York, Paris, Seoul, Beijing, Berlin and Madrid. The size of each system varies in size, as does the power source, depth and speed of the trains, however the systems tend to be almost unanimously popular as a choice of travel mode for commuters.

The Karanasiou et al. (2014) review found 12 studies that had measured exposure in subway systems. Particulate levels were found to be generally highest on the platforms of the system, compared to inside the vehicles themselves. Exposure was also generally higher during peak

travel times. Mean PM_{10} levels were in the range of 103 to 1030 μg m^{-3}, and $PM_{2.5}$ levels between 59 and 375 μg m^{-3}. The highest levels were found in London, Stockholm and Rome. Variation between and within systems was thought to be explained by different brakes types, air-con or natural ventilation systems, different types of rails, and the numbers of trains. A summary table from the review is shown in Figure 2.27 (from Karanasiou et al. (2014)).

PM exposure	Parameter	Equipment	N of samples	Mean, μg/m^3	SD	Min–max
Milan metro, Italy (Colombi et al., 2013)	PM_{10}	Low volume sampler	3 lines tested	103–184 (platform, weekdays)		58–299
Barcelona metro, Spain (Querol et al., 2012)	PM_{10}	High volume samplers, MCV	1 old line and 1 new line tested, 20 days[a]	367 (old line platform)		
				193 (new line platform)		
	$PM_{2.5}$			133 (old line platform)		
				59 (new line platform)		
	PM_{10}	Aerosol monitors, GRIMM and DustTrak	3 days	79 (old line, inside train)		
				45 (new line, inside train)		
	$PM_{2.5}$			25 (old line, inside train)		
				14 (new line, inside train)		
	PM_1			24 (old line, inside train)		
				5 (new line, inside train)		
Budapest metro, Hungary (Salma et al., 2007)	PM_{10}	Tapered element oscillating microbalance, TEOM		155	55	
Paris metro, France (Raut et al., 2009)	PM_{10}	Tapered element oscillating microbalance, TEOM		183 (weekdays)		
				84 (weekends)		
	$PM_{2.5}$			58 (weekdays)		
				29 (weekends)		
(Ripanucci et al., 2006)	PM_{10}	Low volume sampler		381 (platform)		
Prague, metro (Braniš et al., 2006)	PM_{10}	Aerosol monitor, DustTrak	108 journeys	102 (vestibule)		
				113 (inside train)		
Helsinki metro, Finland (Aarnio et al., 2005)	$PM_{2.5}$	Optical monitor, Eberline	12 days	52 (inside the station 4 m above platform level during weekdays from 6 to 18 h)		12–103
	$PM_{2.5}$	Low volume sampler		21 (inside train)		17–45
London metro, UK (Seaton et al., 2005)	PM_{10}	Aerosol monitor, DustTrak	3 days	1037 (platform)		
	$PM_{2.5}$	Aerosol monitor, DustTrak	3 days	343 (platform)		
				170 (driver's cabin)		
Stockholm metro, Sweden (Johansson and Johansson, 2003)	PM_{10}	Tapered element oscillating microbalance, TEOM	257 h	469 (weekdays)		212–722
				336 (weekends)		3.5–280
	$PM_{2.5}$		257 h	258 (weekdays)		105–388
				185 (weekends)		4–69
London, UK (Adams et al., 2001)	$PM_{2.5}$	High flow personal sampler	56	202.3		12.2–371.2
London, UK (Pfeifer et al., 1999)	$PM_{2.5}$	SKC Inc. pump	16 h	246	52	

Figure 2.27: Summary of underground subway studies

Focusing specifically on London, the work of Adams in the late 1990's /early 2000's is oft cited and well known in this area of transport exposure due to it being the first comprehensive particulate matter multi–mode transport user exposure assessment study in the UK. They measured a large number of journeys (465) across many modes (bicycle, car, bus, tube) over different periods in July 1999 ('summer') and February 2000 ('winter') finding mean $PM_{2.5}$ values for the London Underground of 247.2 μg m^{-3} in summer and 157.3 μg m^{-3} for the winter (or an overall mean of 202 μg m^{-3} as shown in the Figure 2.27). Shortly after this publication another paper was published by the same group which tried to apportion the PM to sources (Adams et al. (2001b)). They concluded that most of the PM came from metals from sources such as braking, tyre wear and the tracks. A high iron content was found. They noted that the PM in the underground should not be directly compared with PM above ground as it's composition is very different. In terms of the actual levels measured in the first paper however extrapolation of these measures across the entire underground network including within cabins and on the platforms is not proven. Seaton et al. (2005) found large variations between PM in cabins and on the platforms themselves. In their research (funded by the London Underground to look at at occupational health exposure to $PM_{2.5}$) platform concentrations were 270-480μg m^{-3} and cab concentrations were 130-200 μg m^{-3}. Similarly high to the other studies of the London Underground. This paper also found a high iron content in the PM of 67%. Seaton summarises however that although the concentrations are high, they are unlikely to represent a health risk to workers due to daily time-weighted exposure calculations. Loxham et al. (2013) disagrees. Their research was not included in the Karanasiou review (presumably due to it being published only shortly before the review). They looked at PM_{10}, $PM_{2.5}$ and $PM_{0.1}$ composition finding similar to the other research, that the particles are mostly from interaction between wheels, rails, and brakes with a high iron content. However they suggest that the potential health effects of exposure to the ultrafine fraction of underground PM warrants further investigation, as a consequence of its greater surface area/volume ratio and high metal content.

To summarise, exposure and the related health effects of poor air quality in subway systems has not been extensively studied in most cities. Certainly not to the point where researchers can confidently say what levels of particulate matter the public are being exposed to at different times of the day, in different places of the underground, and what composition that has (and how toxic to the body it is or is not). What can be fairly confidently said however is that concentrations in most systems, and certainly in London, are higher than street-level, in-vehicle, cycling, buses or most other studied transport micro-environments. Although the composition is very different and may be more or less toxic. Due to the number of people that use this transport mode daily (1.171 billion journeys annually in London according to Transport for London (2014b)) for extended periods of time, understanding and being able

to quantify exposure during it is necessary to better estimate people's daily exposure across a range of micro-environments. Certainly they vary hugely from concentrations measured at a subjects place of residence, as is used in the static exposure studies considered earlier.

2.4.3.6 Transport exposure summary

Studies of exposure while people are in-transit between locations using transport modes such as cars, trains, subway systems and bicycles shows a large range of concentration values. There are few exposure studies which account for these variations (alongside outdoor indoor infiltration) to provide exposure data on the daily lives of large numbers of subjects suitable for epidemiological analysis. Until more expansive exposure studies that follow large groups of people of varying time-activity patterns are completed, the ability to discern the range of commute-times specific contribution to total exposure is constrained (Knibbs et al. (2011)).

2.4.4 Dynamic and Hybrid exposure models

Section 2.3 (Static exposure studies) concluded that taking fixed-point or fixed-area measurements, often at one point in space or time, was insufficient to accurately quantify human exposure to air pollution. Exposure varies in space and time due to an individuals activities, the time of day, and the location of the subject. Poor air quality an individual is exposed to will vary depending on how long they spend in public transport each day (and which mode of transport), how long they spend at home or at work, and the concentrations within those environments themselves (Özkaynak et al. (2013)). Therefore one approach would be to deploy personal monitors to large numbers of people, then aggregate and analyse the data to better understand how exposure varies between people and environments.

Section 2.4.1 (Personal monitoring) considered how personal monitoring devices can be used to do this; to better understand individual-level exposure to poor air quality and the drivers thereof. However it concluded by discussing how collecting large enough quantities of accurate data using this method is extremely difficult to do. The following Sections of 2.4.2 (Infiltration) and 2.4.3 (Transport) looked at modelling approaches which enable researchers to quantify exposure to populations in those specific environments (and some discussed the health implications of those environments), but they were not 'joined-up' so as to enable estimates of individuals total daily/annual exposure alongside other micro-environments and times of the day/year.

This section now looks at a small number of research studies which are also modelling approaches, but that try to be more holistic in their exposure assessments, including many micro-environments as well as temporal and spatial variation of the subjects and air quality as inputs.

Models are of course an abstraction of the real world, and as such will almost certainly never represent it with 100% accuracy, but using these methods it is possible to simulate the movements and exposure of a much larger number of subjects than would be possible otherwise. By being aware of the inaccuracies and drawbacks of the model and model inputs, it is also possible to quantify the error and propagation of error, and take this into account for any subsequent health impact studies. For example one key input to an air quality exposure model will normally be an air quality model. This can be evaluated against monitoring site data, and the RMSE noted, then calculated alongside other model parameters to understand the overall model uncertainty. This said, an area of the model which would be difficult to quantify in this manner is the spatial aspect. When modelling the movements of a individual it is hard to numerically quantify this in a form that can be included with the other model uncertainties. If an individual takes one route to work, but

the model predicts another, this would be a difficult error to propagate due to the different units being used (metres rather than pollutant concentrations).

This is a new area of research which is not yet established, and the nomenclature is thus still under development, but a trend towards discussion of 'hybrid' or 'dynamic' models has emerged. 'Hybrid' being a description of a combining of multiple approaches of exposure, such as time-activity diaries combined with modelled air quality data and micro-environmental modelling of transport modes, and 'Dynamic' reflecting the improved temporal scale of data available, for both the populations movement and the variation in pollutant concentrations.

The following sections are samples of current publications in this field to provide a 'baseline' for the model that this research describes. Kousa et al. (2002), Dhondt et al. (2012), de Nazelle et al. (2013), Gerharz et al. (2013) and Reis et al. (2018) all attempt, with varying degrees of complexity, to model exposure of individuals over an entire day (rather than using measurements or only considering one micro-environment).

2.4.4.1 Kousa et al

Kousa et al. (2002) was a very early attempt at this sort of model. They evaluated the temporal and spatial exposure of the population of the Helsinki Metropolitan Area in different micro-environments (home, workplace, traffic and other). For time-activity data (15 min resolution) they used 435 diaries from the EXPOLIS study, and they then linked this to modelled NO_2 data from a dispersion model which outputted grids of 500 m by 500 m for the greater Helsinki area, and 50 m by 50 m for the city centre. For indoor environments they took an I/O ratio of 0.76, including when the subjects were recorded as being in transport. The GIS technology and spatial analysis techniques used in this model would have been quite advanced at the time, and therefore modelling the exposure of 8000 people was impressive. Specifically, modelling the air quality with hourly variation and at 50 m resolution. However in-light of contemporary work this model is simplistic in a number of ways including: only describing four micro-environments (and using the same I/O ratio for all of them), that the time-activity diaries are only at 15 minute intervals, that the data only includes people of ages 25-55, and that they only modelled weekdays. It would also have been interesting to see what difference using this model made on exposure assessment i.e. a comparison with monitoring sites or address-points, such as the research by Dhondt et al. (2012).

2.4.4.2 Dhondt et al

Based in Belgium Dhondt et al. (2012) compared 'static' and 'dynamic' models for exposure to NO_2 and O_3, modelling their subject's behaviours and location based on 8800 activity diaries and 12 km location 'zones', then scaling this to give results relevant for 5 million people. For their air quality inputs hourly concentrations were modelled using a variety of nested models, giving higher resolution nearer to roads and in urban areas, and lower resolution outside of these areas. Time in transport in each zone was also included. Maps showing the exposure over 24 hours accumulated by each zone are shown in Figures 2.28 and 2.29.

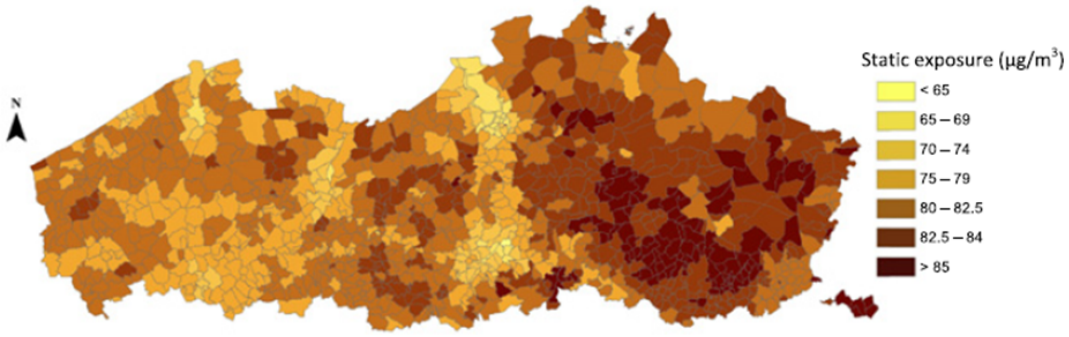


Figure 2.28: NO_2 static exposure results

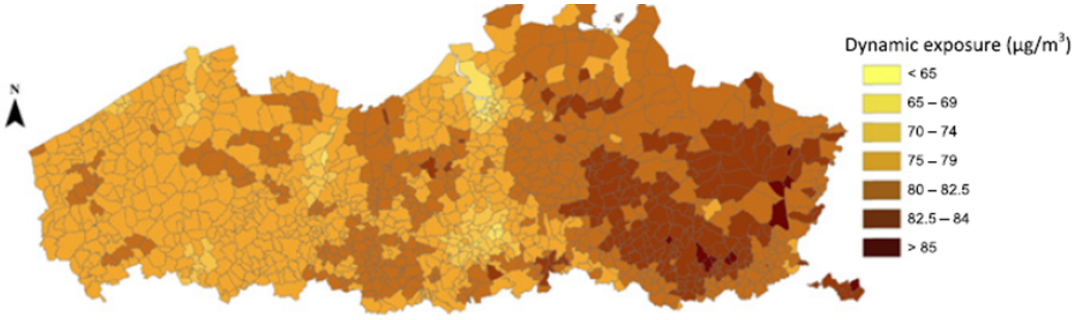


Figure 2.29: NO_2 dynamic exposure results

By using this approach to exposure modelling they found NO_2 was being underestimated by on average 1.2%, and that O_3 was being overestimated by 0.8%. These do not seem particularly large differences, however as an input to a large cohort of subjects in a long time-series analysis they could significantly change any health conclusions. These regional averages also mask much of the variation in individuals, where differences of upto 12% can be found depending on age group, gender and place of residence. Having studied the methods

however, the reliability of the results is open to question. This research is certainly moving towards the type of dynamic exposure model described earlier, as their modelled air quality and time-activity data is of a high quality, however there is no micro-environmental modelling at all, which given the time that we know people spend indoors, must be a large source of uncertainty in the results. As a final note on this publication, the method of grouping the exposure results into time-activity zones rather than by residence of the subject was interesting and informative, showing the areas where greatest miss-classification occurred in a visual and effective manner.

2.4.4.3 de Nazelle et al

The following year de Nazelle et al. (2013) did a similar study for 7 days of 36 subjects in Barcelona, however in addition to using time-activity diaries as an input, the subjects also had software installed on a provided smartphone, which made use of the accelerometer and GPS hardware to record their location and infer their type of activity. After substantial processing of this data, the subjects location and activity was linked with an hourly resolved 5 km x 5 km dispersion model grid of NO_2, and personal exposure was then calculated, including adjustments for six different micro-environments (indoor, bike, bus and tram, car and taxi, metro and train, motorcycle). This study was truly a hybrid study, in that data was collected using personal monitoring (smartphones) but then joined with modelled data and further processing of the data. De Nazelle then calculated the time that the subjects spent in different micro-environments, and compared this to the percentage of their daily exposure in the same micro-environment categories (Figure 2.30).

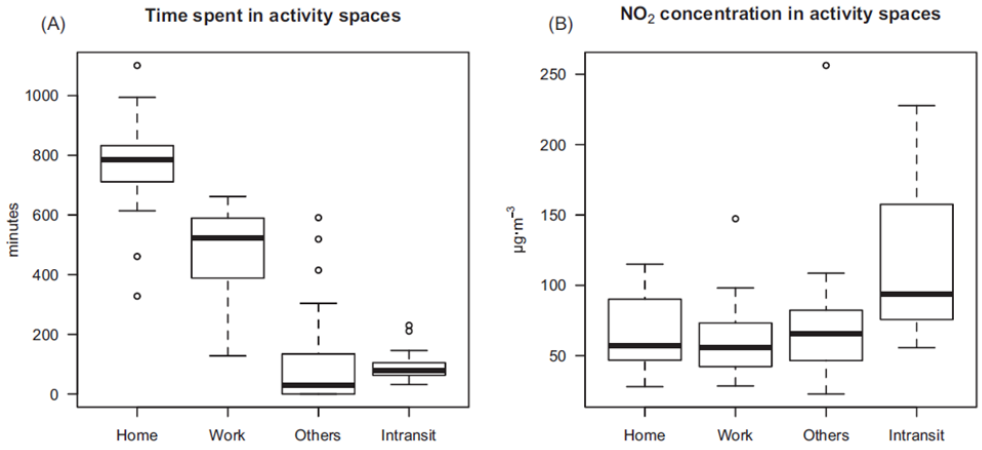


Figure 2.30: Time and NO_2 in activity spaces

From the results we can see that the subjects spent most of their time at home and at work (left), supporting the use of static exposure models where these fixed locations are used,

however when the actual exposure is considered the influence of the time spent in those locations diminishes (right) and 'Others' and 'In-transit' become important (supporting the conclusions of Dons et al. (2011) discussed in Section 2.4.3, 'Transport'). The study concludes by discussing how these techniques could be developed and used for much larger populations (presumably as an input to health studies eventually). However the amount of post-processing of the geographical data, the manual translation of activity diaries to compliment the smartphone data, and the extra battery packs that needed adding to the phones makes this seem less likely. To upscale this to many hundreds or thousands of subjects would need a tremendous amount of coordination and equipment which would need to be distributed, collected and monitored. The data processing would also seem to be difficult without large human resources, and linking to health outcomes would mean the study needed to be designed so that the participants are a statistical representation of the wider study area population, or the numbers surveyed so close to the actual study population that they give a good representation anyway.

2.4.4.4 Gerharz et al

In the same year Gerharz et al. (2013) used similar methods to examine the exposure of 10 people in Munster, Germany. They used activity diaries alongside GPS data again, defined micro-environments using this contextual information of 'home', 'work', 'other indoor', 'transportation', and 'outdoor', joined this to an hourly average PM_{10} dispersion model for the air quality input, and then used I/O ratios for the micro-environments e.g. 1.34 for cars and 2.49 for public transport. The model is summarised in Figure 2.31 below.

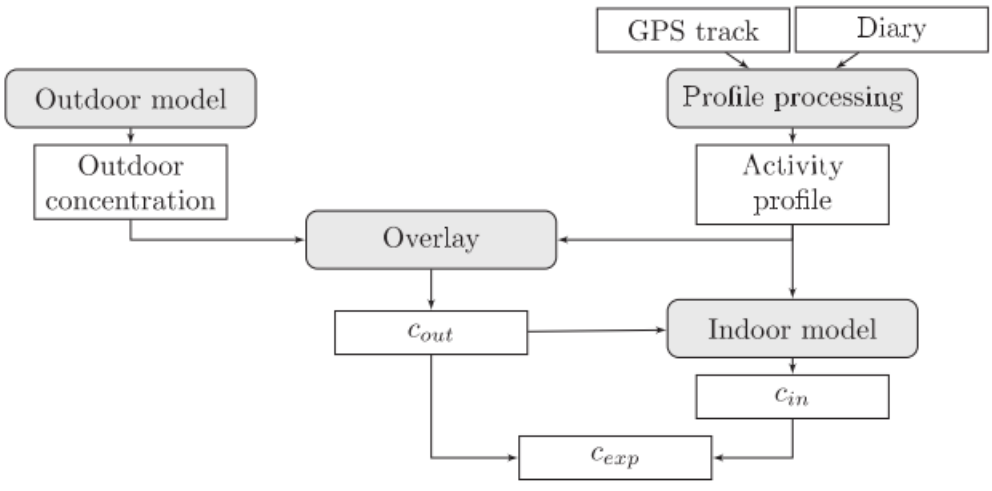


Figure 2.31: The exposure process

The major difference between this and the work of De Nazelle and Dons is that they also gave the subjects personal air quality monitoring data in the form of a Grimm Aerosol Spectrometer. The spectrometer provided information about particle numbers and mass in 32-size classes with a high temporal resolution of 6 seconds. Using this data they were able to evaluate the effectiveness of their modelled exposure by comparing it to real data. The Pearsons correlation coefficient between the mean of modelled and measured data is shown in Figure 2.32.

profile	PM_{10}		
	5 min	1 h	1 h, bgr
p1	0.61	0.81	−0.17
p2	0.49	0.56	0.38
p4	0.41	0.53	0.31
p5	0.59	0.69	0.4
p6	−0.15	−0.28	0.66
p7	0.87	0.94	0.23
p9	0.65	0.67	−0.09
p10	0.11	0.22	−0.73
p12	0.51	0.61	0.5
p13	0.35	0.42	−0.17

Figure 2.32: Pearsons correlation coefficient between the mean of modelled and measured data

There is clearly some variation between the results, with P7 (person 7) being particularly strong, and P10 being particularly weak. Gerharz notes that "Generally, the correlation between model average and measurements is high" which does not seem to tally with the mean of their correlations coming out as 0.44. However they are correct in saying that it does at least "clearly outperform the baseline approach of using urban background measurements as proxy for the personal exposure". Though in terms of using this modelled approach for larger groups of people, the post-processing of GPS data and activity-diaries, general data cleaning, and data linkage would seem to make this an unlikely approach (as with de Nazelle et al. (2013))

2.4.4.5 Reis et al

The most contemporary study to the development of this research is the exposure modelling undertaken by Reis et al. (2018). They combined hourly 1km by 1km estimates of air quality (for the whole of the UK) with 1km by 1km estimates of 'workday' population density v. 'home' population density (based on the UK Census 2011 and the UK Land Cover Map

2015). Calculating exposure for when the population is at home, compared to when they are at home + when they are at work. Given this study modelled the entire UK populations exposure, in a 'dynamic' way, this is a very impressive piece of research; at a much larger scale than any of the previous studies mentioned in this section and with high temporal air quality resolution. This approach has the benefits of being readily transferable to other countries in the world (where similar census and data exist, which are not uncommon), however it could be improved by a) including commute time/mode of transport in the exposure estimates b) including exposure/time spent in indoor microenvironments and c) by modelling movement and air quality at a finer scale than 1km. The model also only allows exposure statistics to be calculated at an aggregate level, making individual exposure estimates difficult, though perhaps example distributions for each 1km grid could somehow be included with further data manipulation.

2.4.4.6 Dynamic and hybrid model reviews

Özkaynak et al. (2013), Meliker and Sloan (2011) and Baxter et al. (2013) have written reviews / position statements on this type of research. Ozkaynak's diagram below (Figure 2.33) is useful for summarising how the complexity of exposure modelling has advanced, and with the complexity the inputs required.

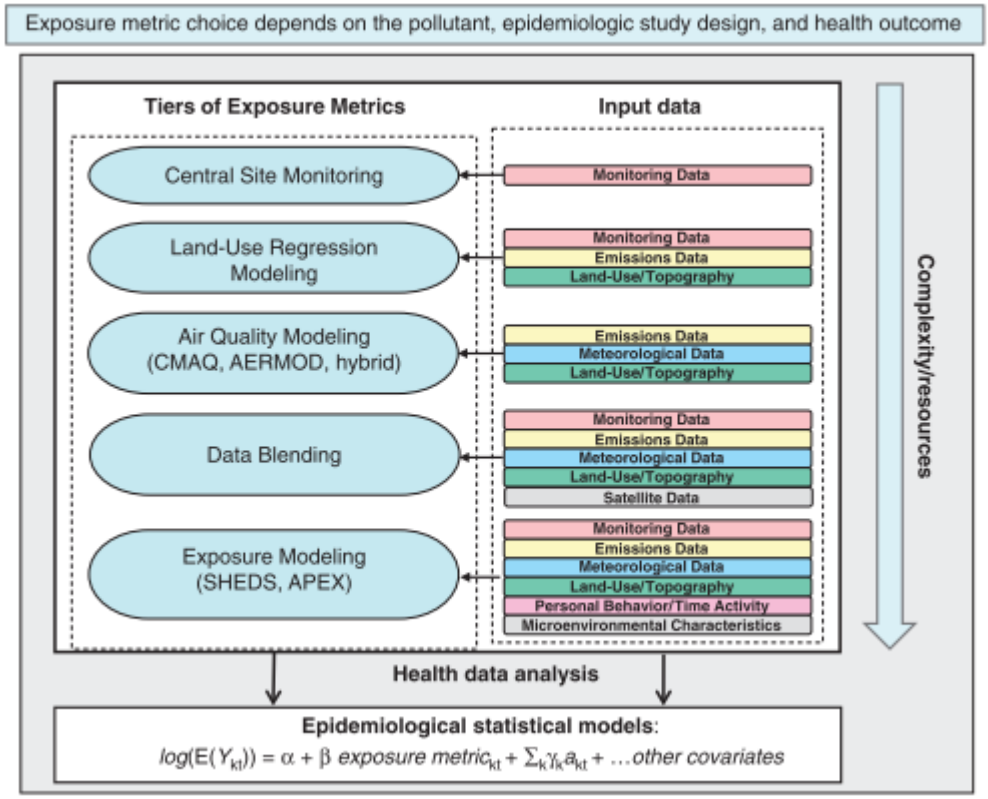


Figure 2.33: The evolution of exposure assessment

They concluded that these new approaches can help refine the significance of air pollution health outcomes, but that this will depend on study-specific characteristics, including epidemiological study design (e.g., time-series vs cohort), the form of of health outcome being considered (e.g., long-term or short-term), which pollutants, and the role of pollutant and building specific indoor infiltration and human activity patterns. Meliker and Sloan (2011) offers a slightly different but useful perspective, by identifying what they consider are the five most important domains to the development of improved spatio-temporal epidemiological models, namely:

1. spatio-temporal epidemiologic theory
2. selection of appropriate spatial scale of analysis
3. choice of spatial/spatio-temporal method for pattern identification
4. individual-level exposure assessment in epidemiologic studies
5. assessment and consideration of locational and attribute uncertainty

Baxter et al. (2013) summarises this emerging area of research well by explaining how, when compared with the use of central-site monitoring data (or other fixed-location methods) the enhanced spatial (and temporal) resolution of air quality or exposure models can impact on resultant health effect estimates, especially for pollutants derived from local sources such as traffic. They recommend that future research develops pollutant-specific infiltration data, improves existing data on human time-activity patterns and exposure to local sources, in order to enhance human exposure modelling estimates. Also that these new approaches are compared with existing approaches to exposure estimation to better characterise estimates in chronic health studies. This research attempt to take this field forward as described.

Figure 2.34 is proposed as a conceptual model of a hybrid/dynamic exposure model, in response to the review of the field by Baxter et al. (2013) and having reviewed the literature in the field during Section 2.4.4. The model should have highly temporal and spatially resolved air quality inputs which consider both indoor and outdoor sources (including regional and local source for the latter), it should be able to model infiltration rates for different modes of transport and building types, it should reflect the multiple micro-environments that people spend their time in (and take account of the temporal resolution of these) and finally it should (for linkage through to epidemiological end-points) be able to consider different breathing rates to quantify exposure and dose for multiple pollutants.

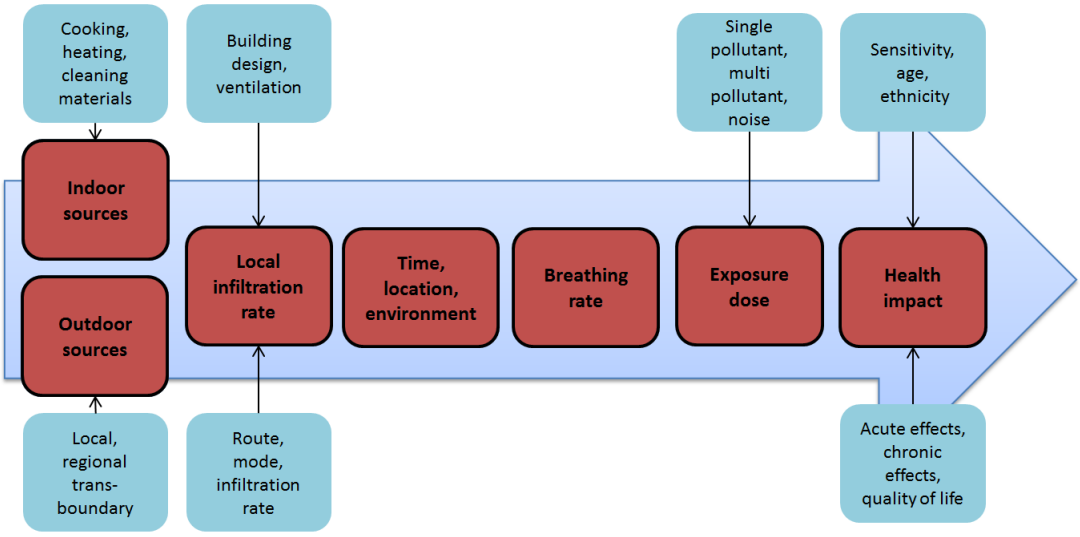


Figure 2.34: A conceptual dynamic exposure model

2.5 Research aims and objectives

Having reviewed literature in the field of dynamic or hybrid air quality exposure modelling, and the general background areas of air quality and health, with the above model in mind, the hypothesis this research proposes is that "*Individual and population level air pollutant exposure can be estimated using time-activity surveys, GIS and routing tools, coupled with high resolution spatio-temporal air quality models, facilitating a greater understanding of the exposure to air pollution in an urban environment*". The research was structured around four primary aims as follows:

1. Reconstruct the time-space activity of London's population
2. Link modelled air quality, estimate exposure, and compare with traditional methods
3. Refine the models estimates of London Underground exposure
4. Evaluate the results

Exposure on the London Underground was focused on in the third results chapter (as oppose to exposure in other environments) due to it becoming apparent during chapter two that the concentrations in this environment are some of the highest that Londoners are exposed to during their typical days.

In more detail the movements of the subject's in the TfL dataset will be modelled, their days reconstructed on a fine temporal and spatial scale, and this data used in conjunction with a novel exposure model. The air quality input used was the CMAQ-UK model (see Section

4.3.1). The results of the model allowed interrogation of exposure by individual, or grouped by various demographics, which enable epidemiologists to increase their understanding of exposure miss-classification. Results are compared to static modelling approaches of the type discussed in Section 2.3. The model is then refined with personal monitoring on the London Underground, before an introduction to validation/testing of the results is presented. The objectives follow the work-plan below.

2.5.1 Modelling Londoners movements

Aim Create a model of Londoners daily movements based upon freely available TfL datasets

Objectives

1. Source, explore and identify key TfL flow datasets
2. Clean and import data to working environment
3. Complete any required modal specific routing (interrogation, querying, storage)
4. Quality assurance / quality check (QA/QC)
5. Analysis by demographic / geographical area

2.5.2 Dynamic exposure modelling

Aim Model exposure to $PM_{2.5}$ and NO_2 of the LTDS-X subjects, and compare with traditional exposure methods.

Objectives

1. Link population movement data to air quality data
2. Incorporate I/O ratios and micro-environmental factors
3. Create a postcode comparison dataset
4. Create a address-point comparison dataset
5. Create a monitoring site comparison dataset
6. Analysis

Thus far, these aims cover the first half of the conceptual exposure model shown in Figure 2.34 (which is repeated here for convenience).

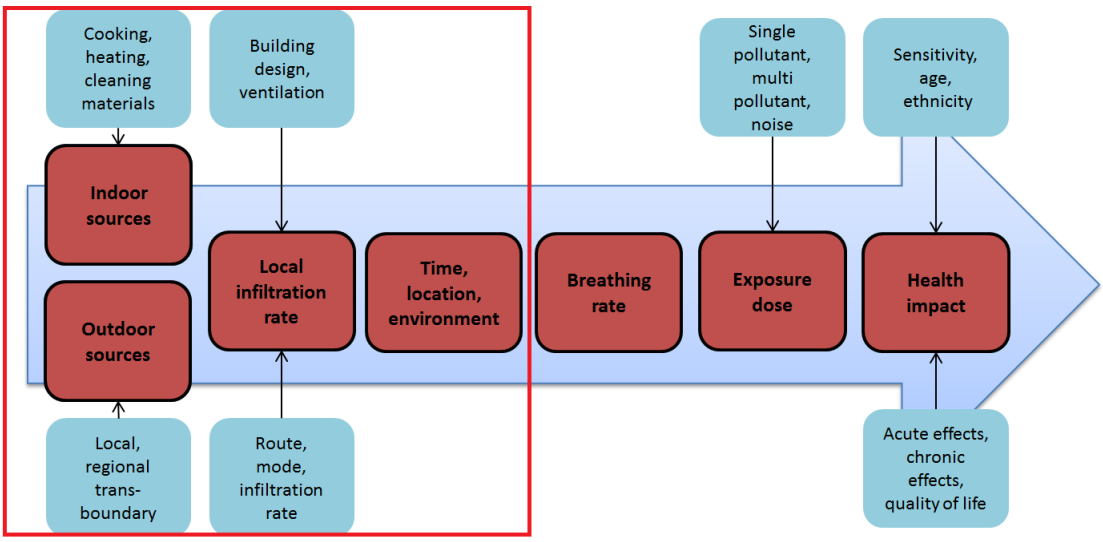


Figure 2.35: A conceptual dynamic exposure model

The following two sets of aims and objectives focus on improving exposure estimates, and evaluating them respectively.

2.5.3 Exposure to $PM_{2.5}$ on the London Underground

Aim Create an exposure model for exposure to $PM_{2.5}$ on the London Underground

Objectives

1. Measure $PM_{2.5}$ across the London Underground network
2. Link measured air quality data to noted time-location data
3. Import, process and clean other datasets for linking i.e. platform depths, station locations.
4. Analysis

2.5.4 Evaluating dynamic exposure models

Aim Develop an understanding of methods to evaluate predictions of exposure from hybrid-type models

Objectives

1. Develop a data collection plan based on simulated and measured datasets
2. Undertake mobile monitoring to collect data representative journey(s)

3. Model exposure of the same journey(s)

4. Analysis: compare the monitored and modelled exposures

3. Modelling Londoners movements

3.1 Aim

Create a model of Londoners daily movements based upon freely available TfL datasets

3.2 Objectives

- Source, explore and identify key TfL flow datasets
- Clean and import data to working environment
- Complete any required modal specific routing (interrogation, querying, storage)
- Quality assurance / quality check (QA/QC)
- Analysis by demographic / geographical area

3.3 Background

As discussed in detail in Section 2.3 (Static exposure & health studies) many air quality exposure studies do not consider the movements of the subjects and/or to varying degrees other factors such as the temporal fluctuations in pollutant concentrations and levels of infiltration into micro-environments such as the home. The aim of this chapter was to process and characterise the London Transport Demand Survey (The 'LTDS', introduced in Section 3.3.1 below) as an input to a hybrid exposure model. This was achieved by estimating the time-space location of the population of London on a minute-by-minute basis using the LTDS dataset, allowing interrogation of the data and such questions to be answered as; how much time do people spend indoors each day, and is there ethnic bias in the distance that people travel to work? To demonstrate the capacity of the model/-dataset visualisations/maps of peoples routes are created and summary graphs of interesting information.

3.3.1 The London Transport Demand Survey

The LTDS is a survey of households in the London area, covering all London Borough's as well as the area between those Borough's and the M25. It is organised by the Strategic Analysis section of the Planning department at TfL. The survey is an ongoing rolling–survey which started in 2005–2006, surveying approximately 8,000 households per year. It is a key input to the first three stages of the London four-step transport planning model (for London (2018)), illustrated in Figure 3.1 below.

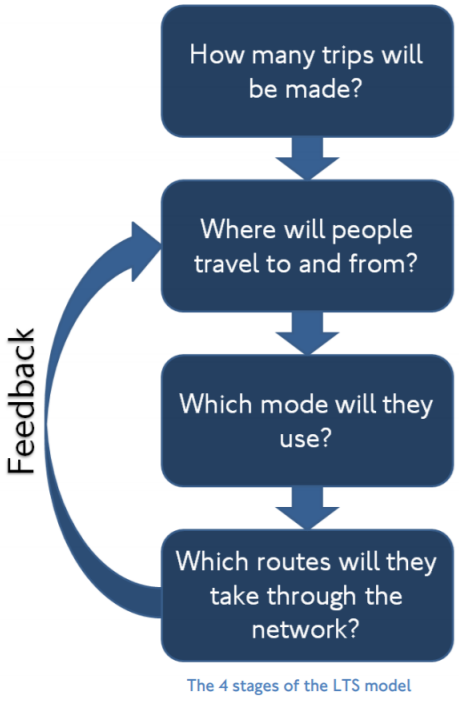


Figure 3.1: The TfL four-step transport planning model

The LTDS captures information on households, the people that live in those households, the trips that those people make in the day, and the vehicles that they use/own. Everyone in the house that is surveyed answers the questionnaires, except for children under the age of 5. There are three questionnaires for each person. The household questionnaire gives details on household structure and includes demographic information such as income, housing tenure and vehicle ownership. The individual questionnaire contains demographic information about the individuals in the household, working status, frequency of transport mode use, driving licenses, public transport tickets held and similar. The third questionnaire is the trip sheet, it captures data on all the trips made on the designated day of travel (which is the same day for all members of the household). The details include the purpose of the trip, the transport modes that were used, the time of the day the trip started, the

time of the day the trip ended, and the origin and destination of the trips. With additional processing, this section of the database can effectively be used to interrogate exactly where each person was during the 24 hours that the survey covers (it covers 4am on the day of the survey, to 4am on the following day, unless that person was in-transit at 4am in which case it continues until the journey is complete).

The LTDS is designed to enable statistically-robust representative estimates of travel patterns and demand in London. Results from a single year are robust enough to analyse at the London-wide level, however combining three or more years data for analysis is preferable. The results can be considered at a London-wide level, or disaggregated to Borough of residence.

The database contains 100 tables of data, however many of these are 'look-up' tables for data in the main tables. For example the year that a household was surveyed is stored in a column called 'hyearid' which simply contains a number between 5 and 9. These numbers allow linkage to a table called YEARID_T where the value 5 can be seen to correspond to the description of '2005/2006'. A summary of the key tables and numbers of records in these tables of the LTDS is shown below (Table 3.1).

Table 3.1: Data contained in the LTDS

Year	Households	People	Trips	Stages
2005–2006	5,008	11,583	29,797	61,542
2006–2007	8,005	18,241	47,029	95,930
2007–2008	7,873	17,926	44,828	91,967
2008–2009	8,134	18,975	43,076	89,701
2009–2010	8,290	19,187	43,475	92,121

3.4 Methods

3.4.1 Data Processing

The LTDS comprises 58 tables of data stored in an Microsoft Access database. The main tables used for this analysis were the 'Household', 'Person', 'Trip' and 'Stage' tables. Table 3.2 below lists some of the more important fields within these tables:

Table 3.2: Key LTDS fields

Table	**Key fields**
Household	hhid: household id hhaboro: Household Borough htdate: Date house survey refers too hincomeei: Household income hhose and hhosn: easting and northing of household householddatatime: This column was created and populated with a full time-stamp
Person	phid: Person id phid: Household id of person ppiwt: Expansion factor of person psexi: Person gender pagei: Person age pegroup: Person ethnic group pegroup: Person ethnic group pseg: Person socio-economic group pnotrips: Number of trips person takes in survey 24 hours
Trip	ttid: Trip id tpid: id of person doing the trip tstagesn: Number of stages in the trip tdpurp: Destination purpose of trip tstime: Trip start time tetime: Trip end time toose and toosn: Trip origin (OSGB36 Easting and Northing) tdose and tdosn: Trip destination (OSGB36 Easting and Northing) tdurn: Duration of trip departuretime: This column was created and populated with a full time–stamp of the trip departure time arrivaltime: This column was created and populated with a full time–stamp of the trip arrival time
Stage	ssid: Stage id stid: Trip id smode5y: Stage mode of transport soose and soosn: Stage origin (OSGB36 Easting and Northing) sdose and sdosn: Stage destination (OSGB36 Easting and Northing) sdurn: Duration of trip (minutes) stagedeparturetime: This column was created and populated with a full time–stamp of the stage departure time stagearrivaltime: This column was created and populated with a full time–stamp of the stage arrival time

Figure 3.2 below shows a basic database schema of seven of the tables as an example of the database structure.

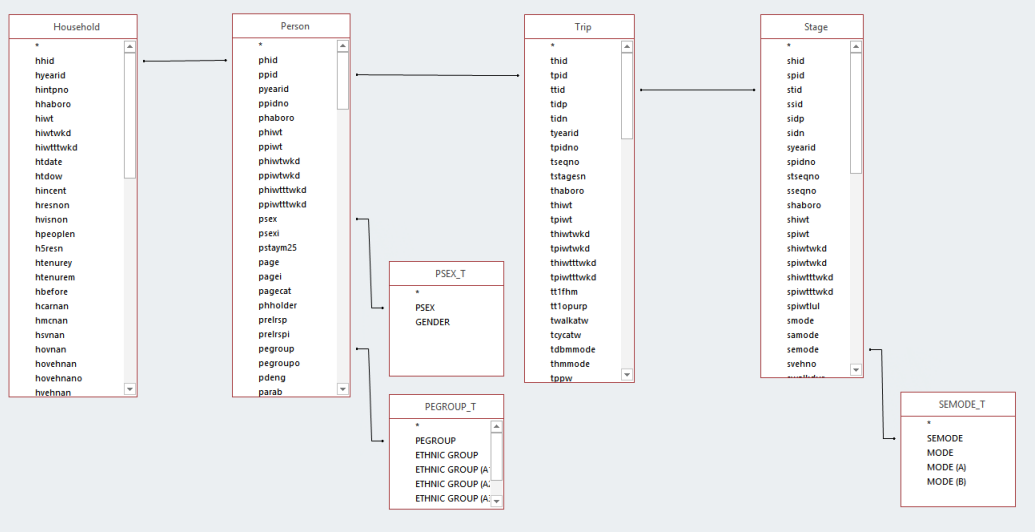


Figure 3.2: MS Access database schema of selection of LTDS tables

The tables listed above were exported as CSV files and then imported into a PostgreSQL installation on a virtual high-spec Ubuntu server. PostGIS was then added to the PostgreSQL installation. SQL scripts were created which added a full time-stamp column to the Household, Trip and Stage tables rather than using the format that the data was stored (dates were stored in the Household table in the form '20060221' (meaning 2006-02-21) and times in the Trips table as '730' (meaning 07:30) which made performing temporal queries on them difficult). Further complications arose in the Trips and Stages tables as the survey period begins and ends at 4am (or once the persons final trip was completed). So times in the Trips table showed as, for instance ,'300' which meant 3am, however that actually meant 3am on the morning of the day after the survey date, rather than the survey date itself. Another factor was that the stages table did not contain any stage departure times or stage arrival times, only the duration of each stage. Therefore the stages table needed cross-referencing against the trip table and then summing incrementally by stage id in order to calculate the correct stage departure and arrival times (and then extensively checking).

3.4.2 Data cleaning

Data cleaning was required throughout each stage, some of which are listed as bullet-points below. An early decision was taken whereby if any data linked to a Person was clearly incorrect, or not recorded properly, then that person was removed from the dataset entirely. For example if a person had a journey which ended 15 minutes after the next journey begun, then that person was removed from the analysis rather than attempt correction of the data.

Removal of records from the dataset in instances like these typically occurred in the following situations:

- Trip start and end times were mis-aligned
- Stage start and end times were mis-aligned
- Stage transport mode was missing or refused
- Key demographic data was missing or refused
- Stage start and end locations were mis-aligned with the next or previous trip by greater than 80 metres
- LTDS respondent did not live in London (the survey stretches just beyond the London Borough boundary)
- Location of stage start or stage end were missing.
- Transport routing for a mode was not available e.g. Bus journeys outside London

To ensure that removal of subjects (9%) from the dataset had not compromised the overall structure and statistical strength of the data (and therefore made using the weighting factors to scale the data unsuitable), t-test and WilcoxonMannWhitney tests were performed to compare the original data, and this subset of data. No statistical significance was found when comparing age, Borough of residence, ethnicity, income, sex, distances travelled each day, or the amount of time in transport.

3.4.3 Mode-specific routing

To calculate the locations of the subjects while they were travelling between places the stages table of the LTDS was used, specifically the start easting and northing, the destination easting and northing, the transport mode, and the stage departure time. These columns were extracted from the database directly into R using the RPostgreSQL extension and stored in an R dataframe.

In the first iteration of this research, routing graphs in the PostgreSQL+PostGIS database were created for the road and London Underground, to undertake estimation of the routes that survey participants took between locations. The data was sourced from OpenStreetMap and TfL, processed, the 'pgRouting' extension to PostgreSQL installed, and routes were generated.

The benefit of this approach to routing was that the network could be easily customised, for example road speed-limits can be adjusted (which will affect routes chosen by the algo-

rithms), a preference for certain road types factored in, one-way restrictions either ignored or obeyed. There was also a certain independence to this approach i.e. no reliance on an external provider keeping the dataset up-to-date or needing a web connection for the routing to take place (as is required for the API approach discussed shortly). However having spent some time doing routing in this manner, we decided to use routing APIs instead. This change was deemed necessary as networks for the other transport modes appeared difficult and time-consuming to set-up, the amount of data to download and process was large and complex (depending on the size and complexity of the graph), and that a new network would be required for each transport mode. This was particularly pertinent for the LTDS as routing on many different transport modes is required including underground, bus, cycling, driving, walking and overground/mainline trains - and data to create networks for each of these modes is difficult to obtain and maintain, particularly for bus routes and train lines as the data is held by the relevant companies that operate these services and even when made publicly available is often not in a suitable format or there is missing information.

APIs offered an alternative approach to routing. Instead of building and maintaining network datasets in our own database/server, the networks are hosted by commercial and non-commercial organisations who allowed interaction with the data (but not to edit it or use it in ways which they do not allow). A user, having read the documentation, first forms a query string (similar to a website address but longer and more complex) and then sends this to the organisations API service. The API interprets the request that the user has made, and returns the data required. In the case of routing,the request from the user is typically a start-point, end-point, transport mode and time of travel, and the return from the API is a route made up of such information as a list of tube stations, or coordinates, or text instructions. The amount of detail that can be submitted as a request, and the amount of detail that is provided by the reply, varies by service. A typical request string to the TfL directions API is shown figure 3.3.

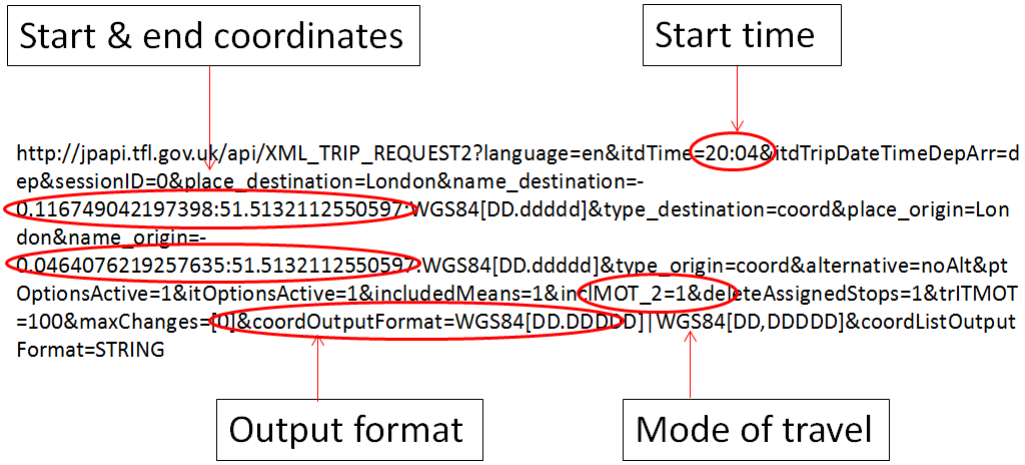


Figure 3.3: A 'call' to the TfL routing API

This 'request' is submitted to the API using software of the users choice, code is written to parse (made into a format that the software can read) the data that is returned, and then the results are stored and used as the user wishes.

For this research a search of routing API services was undertaken based upon the list detailed in the Wikipedia page 'Online Routers' (`http://wiki.openstreetmap.org/wiki/Routing/online_routers`) as well as an internet search of the term 'Online routing API'. A review of those considered is presented in table 3.3. A small handful of other APIs were dismissed before this stage as they were unsuitable for reasons such as geography (e.g. only covers Germany) or the API does not expose the route to the user (e.g. At the time of writing, Bing Maps Directions plots the line onto a map, but does not give the actual data for the line back to the user).

Table 3.3: Summary of suitable routing APIs

Service & coverage	**Available transport modes**	**Usage limits**	**Documentation**	**Simplicity of use**	**Customisation and notes**
Google Directions API (Worldwide)	Car, pedestrian, cycling, public transport	750 per day (without license)	Provided, with many examples	Very simple to use	Limited. Parameters such as use of tolls, language of content, shortest or quickest. Public transport modes cannot be distinguished
OpenRouteService (Worldwide, but dependant on OSM data quality)	Car, pedestrian, cycling	Unlimited	In the form of a Wikipedia entry. Covers some examples but is not comprehensive.	Very simple	Limited. Road types, languages, zoom levels
Project OSRM (Open Source Routing Machine) (Worldwide, but dependant on OSM data quality)	Car	Unlimited	Basic instructions provided on GitHub	Very simple to use	Limited. Zoom level and output file type
Transport for London (Greater London)	Underground, overground, bus, train, tram, boat, cable—car, Docklands Light Railway	Unlimited (once registered)	Extremely comprehensive manual, but unclear in many places and often seemingly contradictory	Very difficult. Only simple examples are provided. So many parameters must be considered for a properly formed request	Fully customisable in almost every way. However it is limited to Greater London and does not alert when routes go outside.
MapQuest (Worldwide)	Car	Unlimited	Basic instructions provided on GitHub	Very simple to use	Limited. Zoom level and output file type

Due to restrictions on usage limits it was decided to use a selection of routing APIs, depending on the transport mode required. Pre-processing was also required to simplify and harmonise the LTDS transport modes before this, for example the LTDS contains transport modes such as 'Car (passenger)', 'Taxi', 'Van (driver)', 'Van (passenger)' and 'Motorcycle (driver)' which were combined to the mode of 'car'. The transport modes and the API used for those are shown below in Table 3.4.

Table 3.4: API used for each LTDS transport mode

LTDS Transport Mode	**API**
Walking	OpenRouteService
Cycling	Google Directions
Train	TfL Journey Planner
Overground	TfL Journey Planner
Underground	TfL Journey Planner
Docklands Light Railway	TfL Journey Planner
Bus	TfL Journey Planner
Car	Project OSRM API

Each journey request was therefore formed into a suitable URL request, similar to the example in Figure 3.3, and then the request was sent to the routing APIs. A small extract of the large XML file that is received in response to a request to the TfL Journey Planner API is shown in Figure 3.4. This was parsed using the RJSONIO and XML packages, and the route between the two locations extracted and stored (and then decoded in the case of the Google API which returns routes as encoded polylines). The route was then formed into a linestring data type (which PostGIS can store as a spatial object) and stored back in the database.

```
<itdCoordinateString decimal="." cs="," ts="&#x20;">
-0.11076,51.48509 -0.11113,51.48538 -0.11121,51.48548 -0.11131,51.48573 -0.11135,51.48625 -0.11136,51.48640
-0.11138,51.48702 -0.11132,51.48770 -0.11132,51.48770 -0.11132,51.48778 -0.11130,51.48792 -0.11127,51.48820
-0.11124,51.48846 -0.11124,51.48862 -0.11123,51.48879 -0.11123,51.48887 -0.11124,51.48900 -0.11121,51.48937
-0.11115,51.49005 -0.11115,51.49005 -0.11109,51.49074 -0.11107,51.49097 -0.11106,51.49115 -0.11103,51.49165
-0.11100,51.49183 -0.11097,51.49234 -0.11094,51.49274 -0.11094,51.49279 -0.11094,51.49279 -0.11094,51.49308
-0.11091,51.49332 -0.11086,51.49400 -0.11094,51.49547 -0.11105,51.49569 -0.11105,51.49569 -0.11121,51.49601
-0.11145,51.49647 -0.11169,51.49695 -0.11172,51.49704 -0.11189,51.49752 -0.11200,51.49778 -0.11217,51.49824
-0.11225,51.49832 -0.11225,51.49832 -0.11234,51.49843 -0.11241,51.49855 -0.11245,51.49868 -0.11248,51.49888
-0.11182,51.49951 -0.11167,51.49968 -0.11167,51.49968 -0.11161,51.49973 -0.11115,51.50023 -0.11064,51.50091
-0.11051,51.50142 -0.11047,51.50159 -0.11045,51.50169 -0.11041,51.50178 -0.11033,51.50185
</itdCoordinateString>
```

Figure 3.4: An example of the route coordinates from a XML response from the TfL API

The cleaned LTDS data contained 45,079 people, who took 98,770 trips during the day of the survey. Each trip consisted of multiple stages that needed routing independently, totally 340,754. These 340,000 routes were routed using the relevant API, and the result stored as a linestring in the PostGIS database linked to the id ('ssid') of the stage.

3.4.4 Quality checking

Spatial cleaning of the routes was now performed. This is discussed earlier in Section 3.4.2 (Data cleaning). It mostly consisted of error-checking of the routes by creating visualisations of the transport modes to manually identify large errors e.g. 'tube' journeys in Birmingham, or taxi journeys across the Irish sea. The results of the underground and Docklands Light Railway (DLR) routing are shown in 3.5 below.

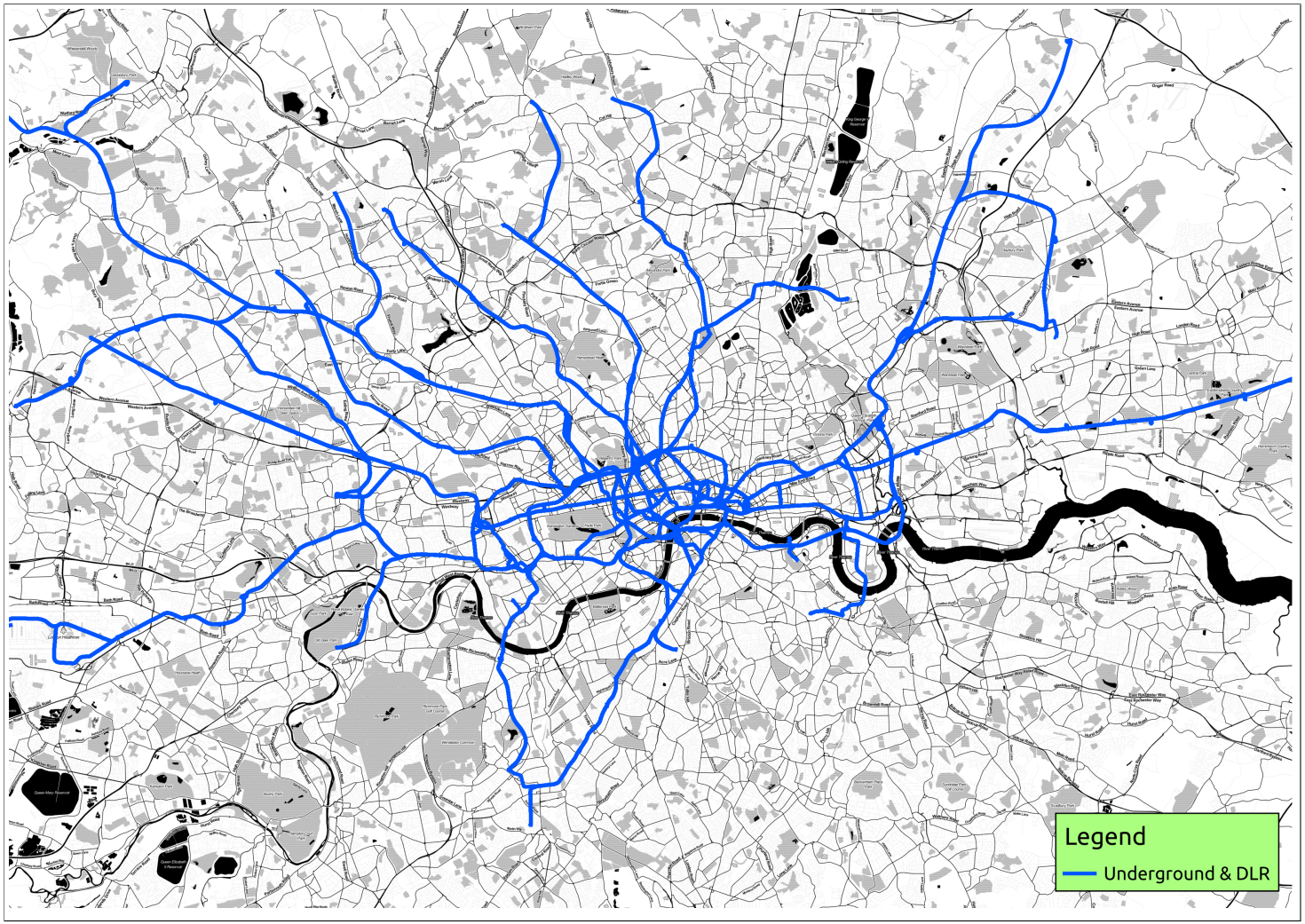


Figure 3.5: A visual check that underground routing results reflect locations of London underground lines

3.4.5 Data manipulation

A 'base table' for each subject in the cleaned and routed LTDS data (45,079) was now created (and named the 'hybrid_location' table). This table contained four columns, person ID (ppid), time (pointtime), mode (mode) and location (thegeom). A blank row for every minute of the subjects 24 hours was created (45,079 subjects, multiplied by 24 hours, multiplied by 60 minutes). The table was then populated with data from the Stage table, by taking the line-strings of each route and splitting them into minute-by-minute interval points using bespoke spatial interpolation SQL scripts, and then matching those points to the correct time in the new table. The mode and id of the stage were also copied over for ease of future reference. A graphic illustrating simple linear spatial interpolation, respecting a time-series, is shown in 3.6 below.

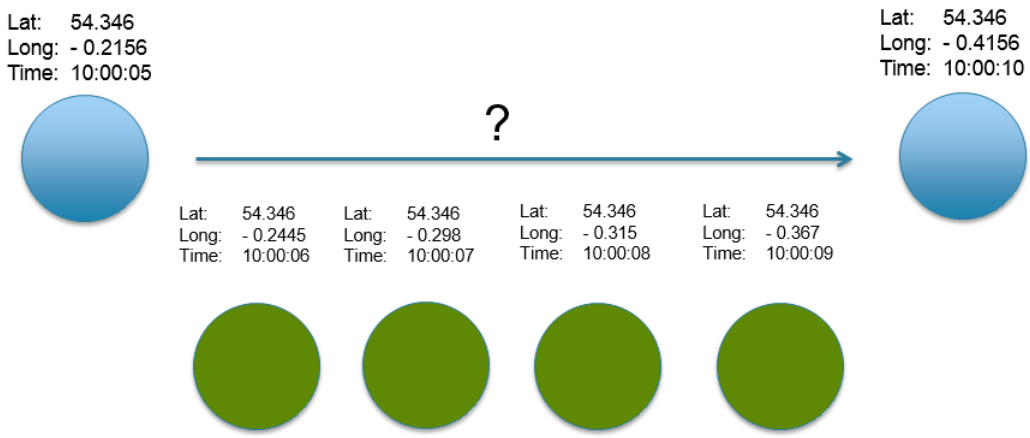


Figure 3.6: The time and location between two known locations and times were calculated using custom-made SQL scripts and the spatial functions of PostGIS

Between trips subjects were presumed to be stationary and indoors at the final point recorded of the previous trip. At the start and end of a day subjects locations were presumed to be the starting point of their first trip, and the ending location of their last trip respectively (and again, indoors). The result of this process was the location of the 45,079 people on a minute-by-minute basis for 24 hours. An extract of the final table is shown below in 3.7 (mode 0 = indoors, 1 = walking, 3 = car).

ppid numeric	point_time timestamp without time	the_geom geometry(Point,27700)	mode numeric	ssid numeric
50500201101	2006-02-21 08:55:00	0101000020346C0000000000	0	
50500201101	2006-02-21 08:56:00	0101000020346C0000000000	0	
50500201101	2006-02-21 08:57:00	0101000020346C0000000000	0	
50500201101	2006-02-21 08:58:00	0101000020346C0000000000	0	
50500201101	2006-02-21 08:59:00	0101000020346C0000000000	0	
50500201101	2006-02-21 09:00:00	0101000020346C0000000000	0	
50500201101	2006-02-21 09:01:00	0101000020346C0000000000	0	
50500201101	2006-02-21 09:02:00	0101000020346C0000000000	0	
50500201101	2006-02-21 09:03:00	0101000020346C0000000000	0	
50500201101	2006-02-21 09:04:00	0101000020346C0000000000	0	
50500201101	2006-02-21 09:05:00	0101000020346C0000000000	1	505002011010102
50500201101	2006-02-21 09:06:00	0101000020346C0000000000	3	505002011010102
50500201101	2006-02-21 09:07:00	0101000020346C0000000000	3	505002011010102
50500201101	2006-02-21 09:08:00	0101000020346C0000000000	3	505002011010102
50500201101	2006-02-21 09:09:00	0101000020346C0000000000	3	505002011010102
50500201101	2006-02-21 09:10:00	0101000020346C0000000000	3	505002011010102
50500201101	2006-02-21 09:11:00	0101000020346C0000000000	3	505002011010102
50500201101	2006-02-21 09:12:00	0101000020346C0000000000	3	505002011010102

Figure 3.7: An example of data and structure in the hybrid location table

3.5 Results

The hybrid location table was used to examine the movements of the subjects of the LTDS on a fine temporal and spatial scale. The data was explored to better understand the detail available, to check that it was suitable for use as the basis of a dynamic exposure model, and to see whether there were any interesting results already present before moving on to exposure modelling.

3.5.1 Visual inspection of individuals

Before starting to analyse the results, Figures 3.8 and 3.9 were created and show the results of two randomly selected individuals from the dataset. This was done as another quality check. The first (3.8) shows a fairly simple reconstructed day. The Person made two car journeys around the local neighbourhood, as well as one short walking trip. Other than visiting those three local locations, they spent their time at home. The second (3.9) shows a much more complicated picture. The individual lives near Norbiton. In the morning they walked to Norbiton station, took a train to Waterloo, and then caught a bus to Clerkenwell (presumably for work). At lunchtime they walked around the local neighbourhood (buying lunch?). At 20:13 they took a taxi back to Waterloo, where they got a train back to Norbiton, and walked home from the station to their house.

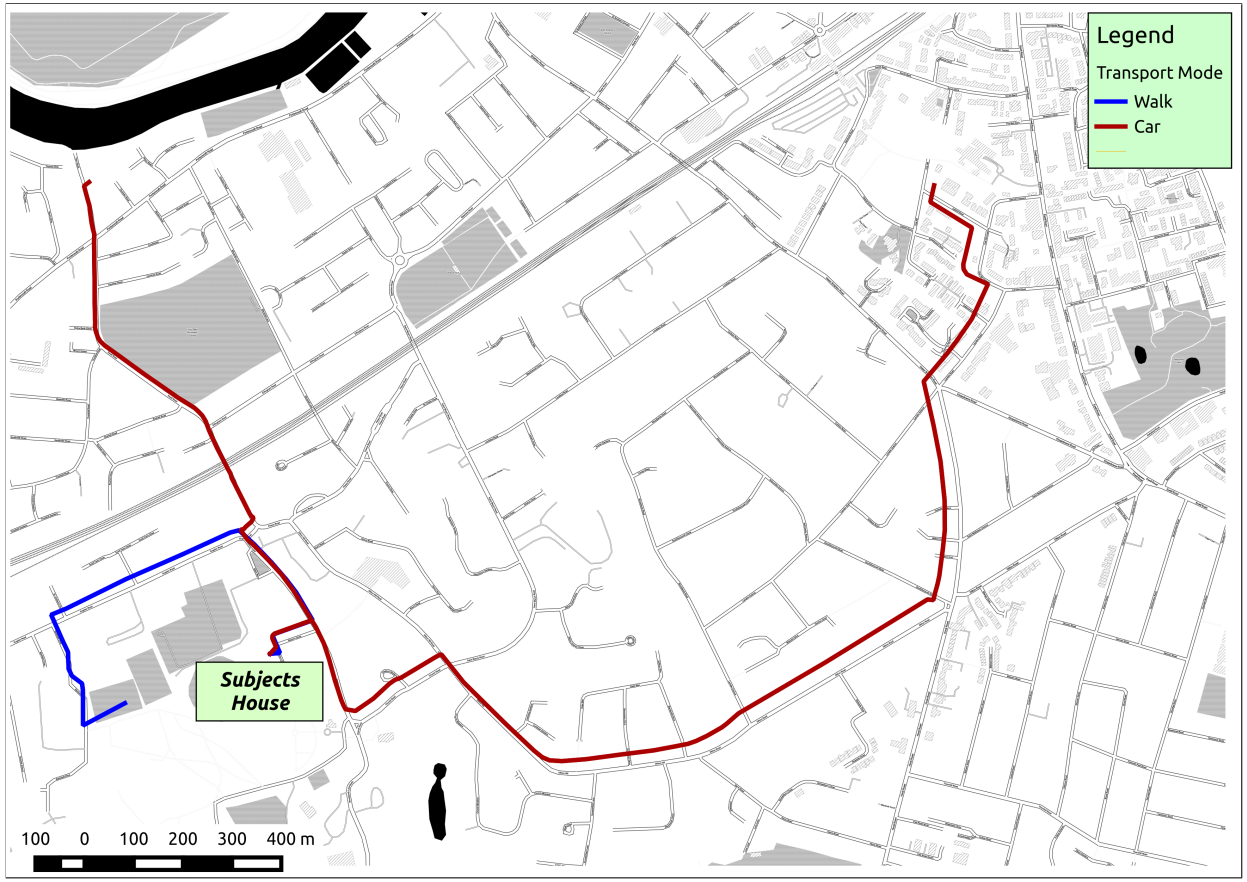


Figure 3.8: Example one of the estimated movements of a subjects day

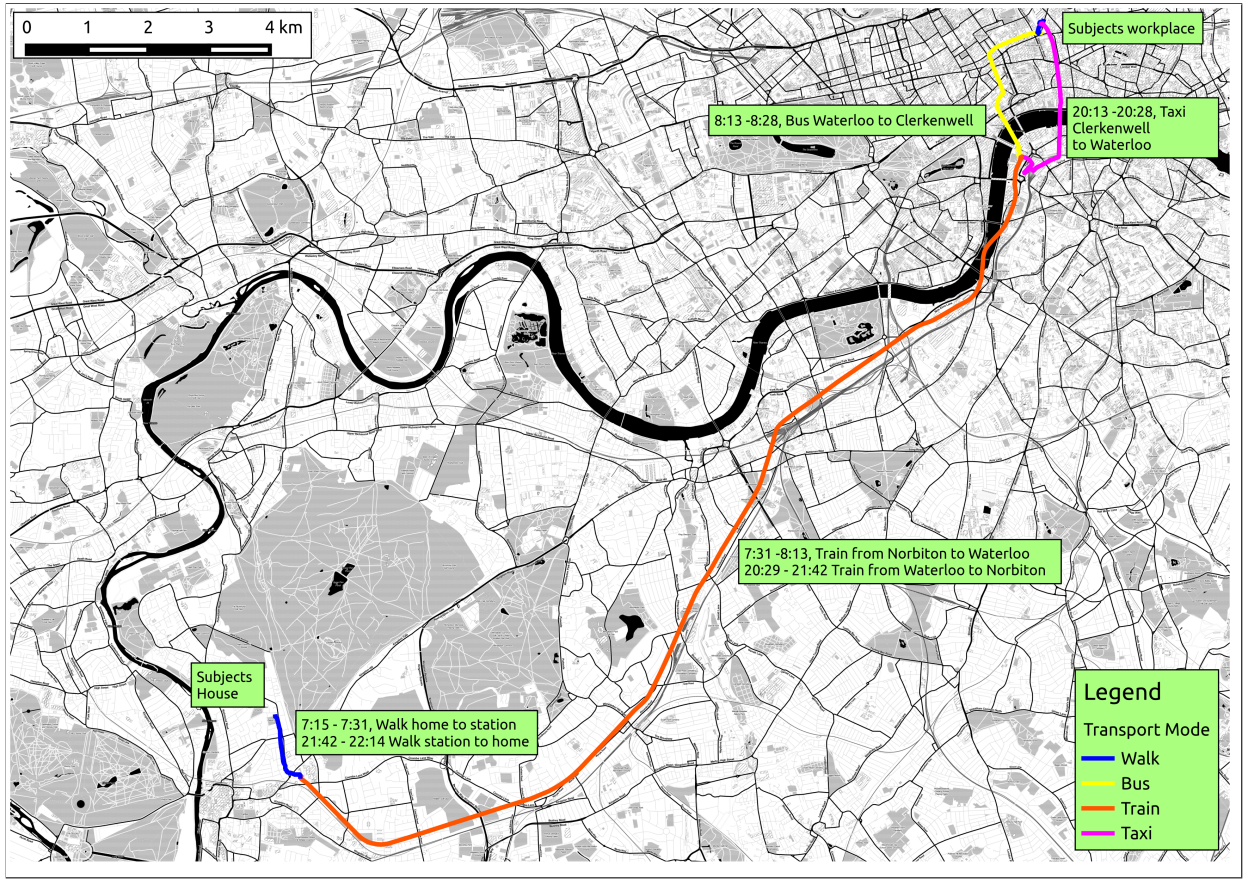


Figure 3.9: Example two of the estimated movements of a subjects day

3.5.2 Journey start and end times

By grouping the start and end times of the stages by hour, and then summarising by the number of stages within that hour, a histogram of journey start times and end times was created (Figures 3.10 and 3.11).

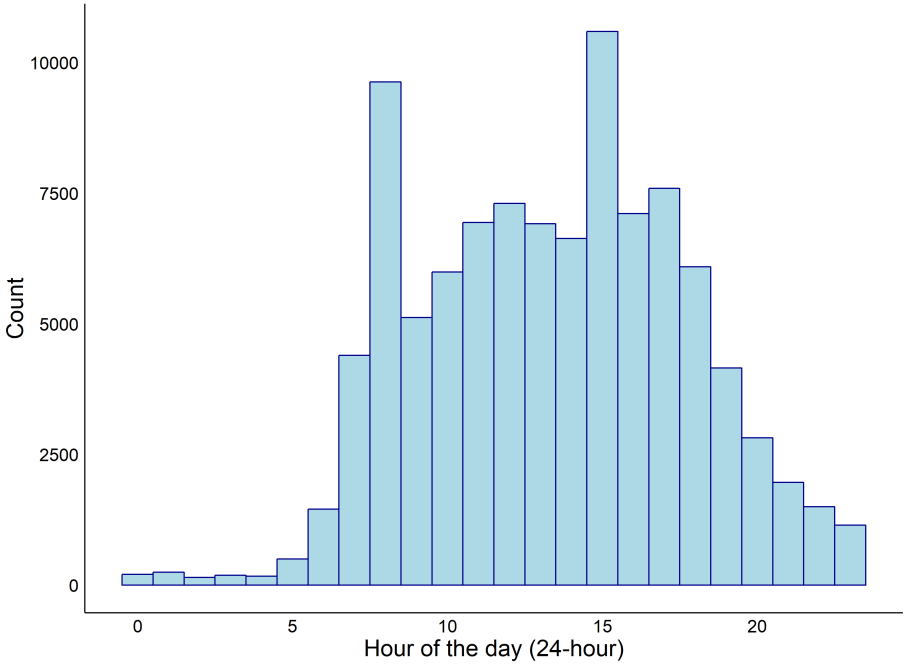


Figure 3.10: Histogram of when the population of London start trips

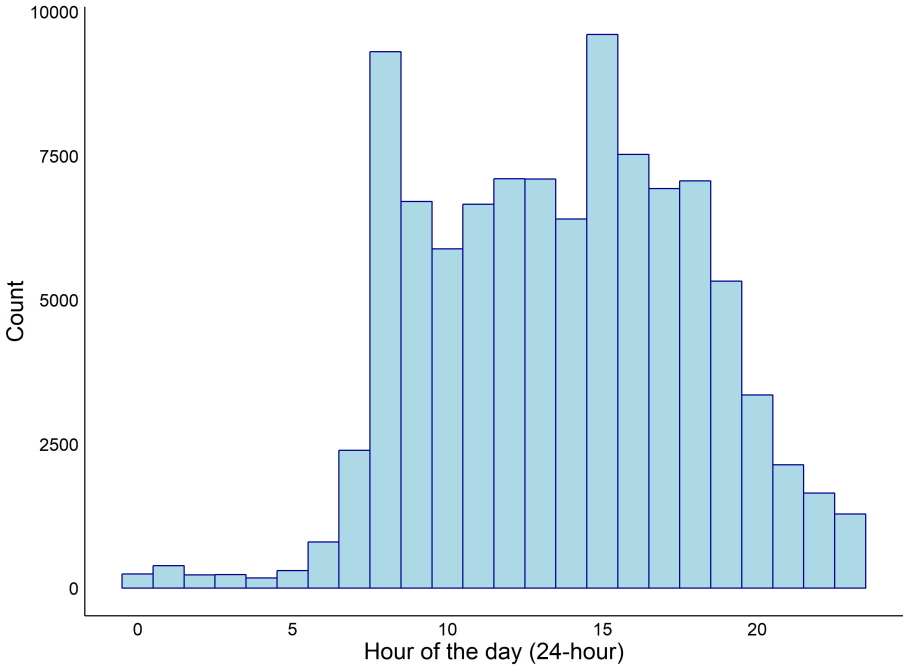


Figure 3.11: Histogram of when the population of London end trips

Figures 3.10 and 3.11 above show peaks of travel around times that might be expected (morning and evening 'rush hours'), however it is interesting to note that considerable travel occurs during the day (although at this stage the data is not split by specific days, so this may be influenced by travel on a Saturday and Sunday). It was also interesting to see how trips during the evening (4pm-8pm) are much more dispersed than in the morning, suggesting that people tend to leave work over a longer time-frame than when they start work.

3.5.3 Journey distances by gender

By summing the total distance travelled by each person over 24 hours, and then grouping by gender, graph 3.12 shows differences in total travel distance.

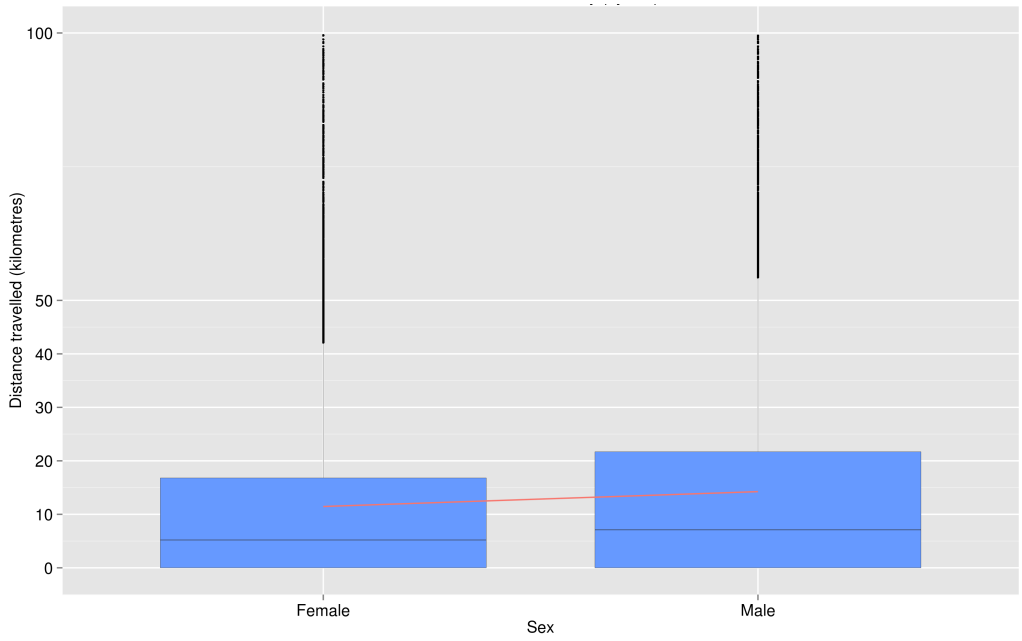


Figure 3.12: Boxplot of distances travelled, by gender (outliers >100 km omitted for clarity, red line links each mean)

Men had a mean travel distance of 18.28 km, and Females a mean of 13.89 km. Whether this is because women tend to stay at home and care for children more than men, or whether the journeys that men are required to do during their day take them further was not clear (but could be investigated further as the dataset contains a great deal of contextual information).

3.5.4 Journey distances by income

A similar method was then used but the demographic of household income was used as the variable to be considered rather than gender (individual-level income variables were unfortunately not collected). The result is shown in Figure 3.13.

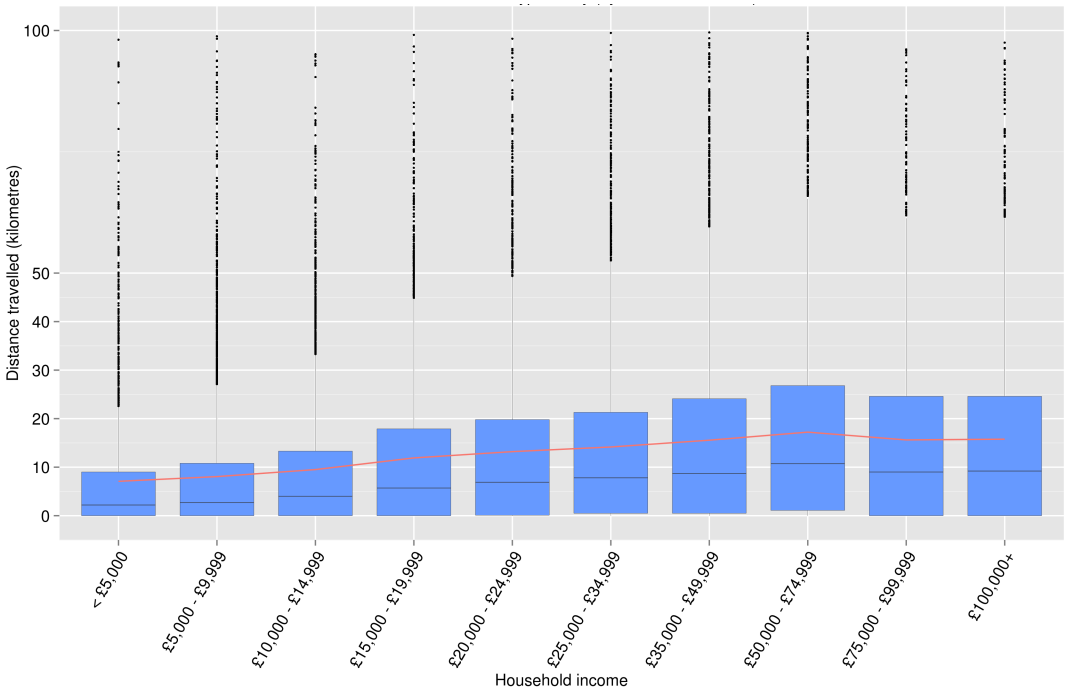


Figure 3.13: Boxplot of distances travelled by income group (outliers >100 km omitted for clarity, red line links each mean)

Interestingly the data shows that subjects that live in households with a higher income travel further than households with a lower income level (albeit with a levelling off and even slight dip around £75,000). Perhaps reflecting that lower income households tend to work in less-skilled jobs that are more locally available.

3.5.5 Journey distances by age group

Distance of travel by age group was now plotted in Figure 3.14.

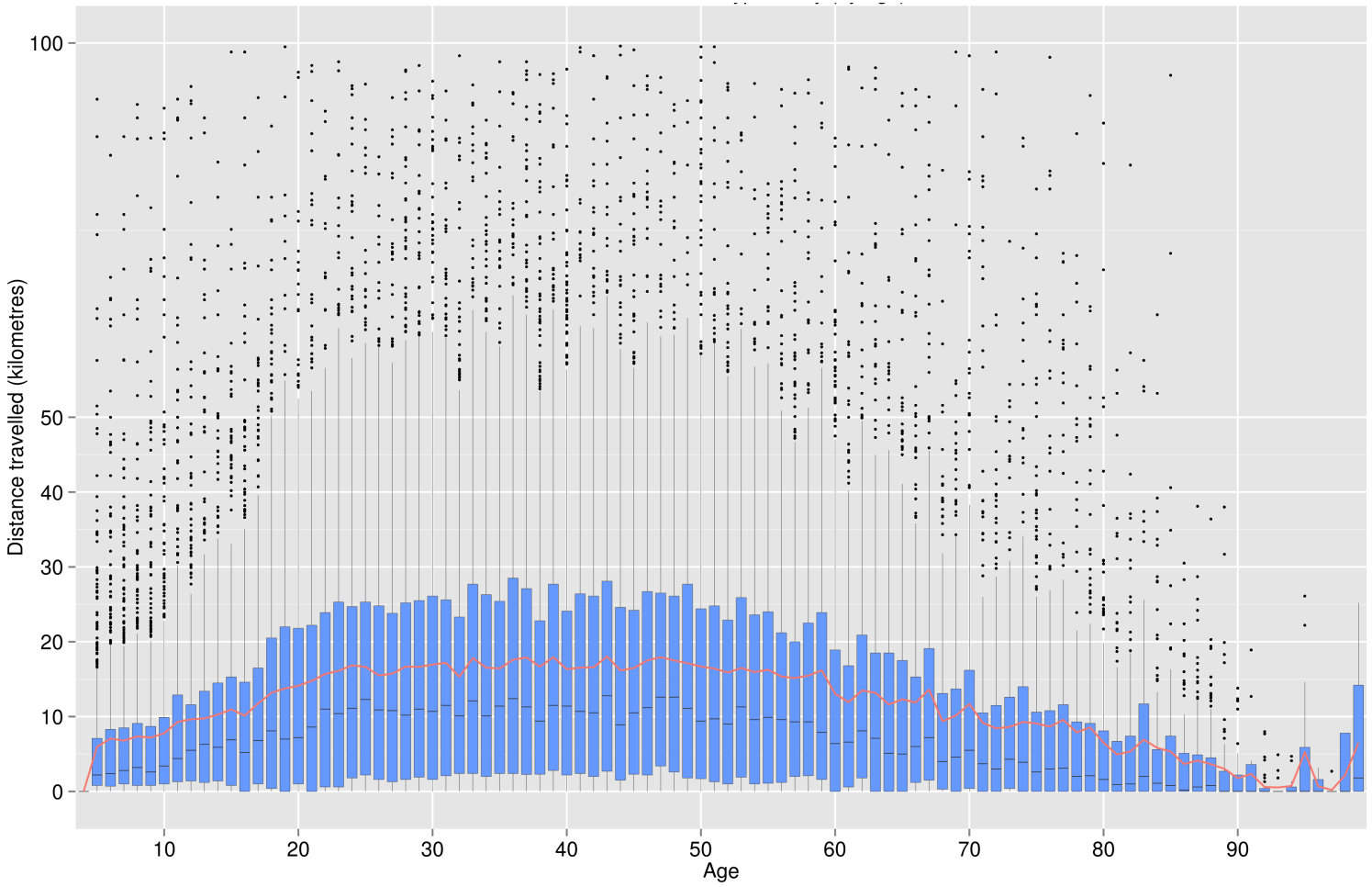


Figure 3.14: Mean distances travelled by age group (outliers >100 km omitted for clarity, red line links each mean)

Figure 3.14 showed a clear rise in the distance that people travel each day as they get into their late teens, which then becomes fairly steady until mid-50s at which point the distances start to decline again. The gradient of the slope between 60 and 90 is much more steady that between 10 and 30 at the other end of the age range, perhaps showing that the ages that people retire are more dispersed than the ages at which people start work.

3.5.6 Journey distances by Borough of residence

The distances that the people of London travel, depending on their Borough of residence, was now considered. The centre of London was defined (the monument outside of Charing Cross station) and then the distance between this point and the centroid of each London Borough was calculated (Figure 3.15). Figure 3.16 was then created, with the Boroughs ordered by this distance metric i.e. the centroid of the City of London is the closest Borough centroid to Charing Cross, so was plotted first. This ordering was undertaken to test the hypothesis that individuals living nearer the centre of London would spend less time travelling than those in outer London.

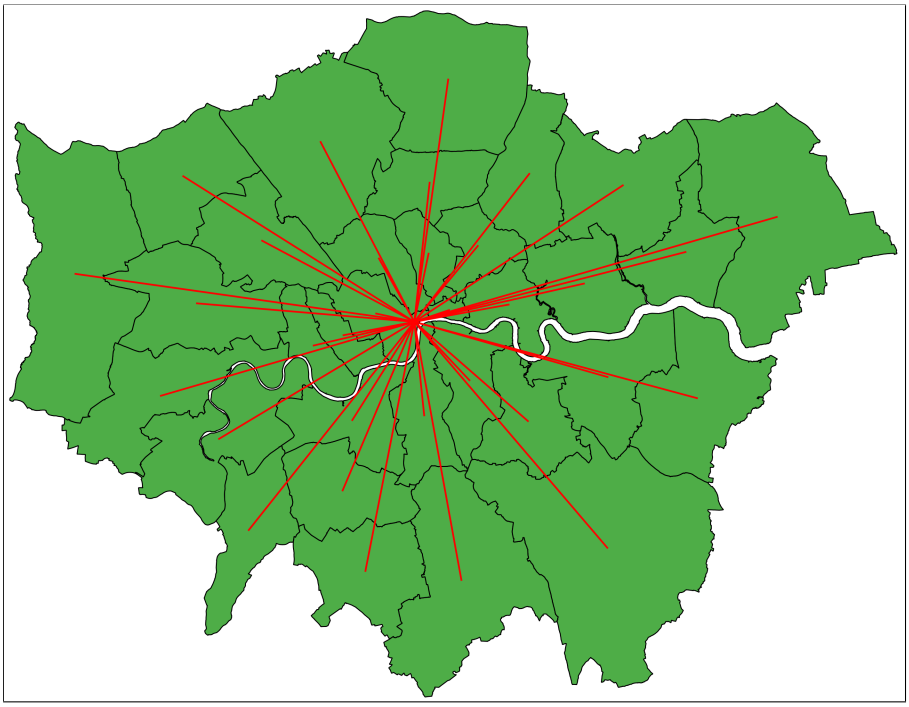


Figure 3.15: Calculating Borough centroids to Charing Cross

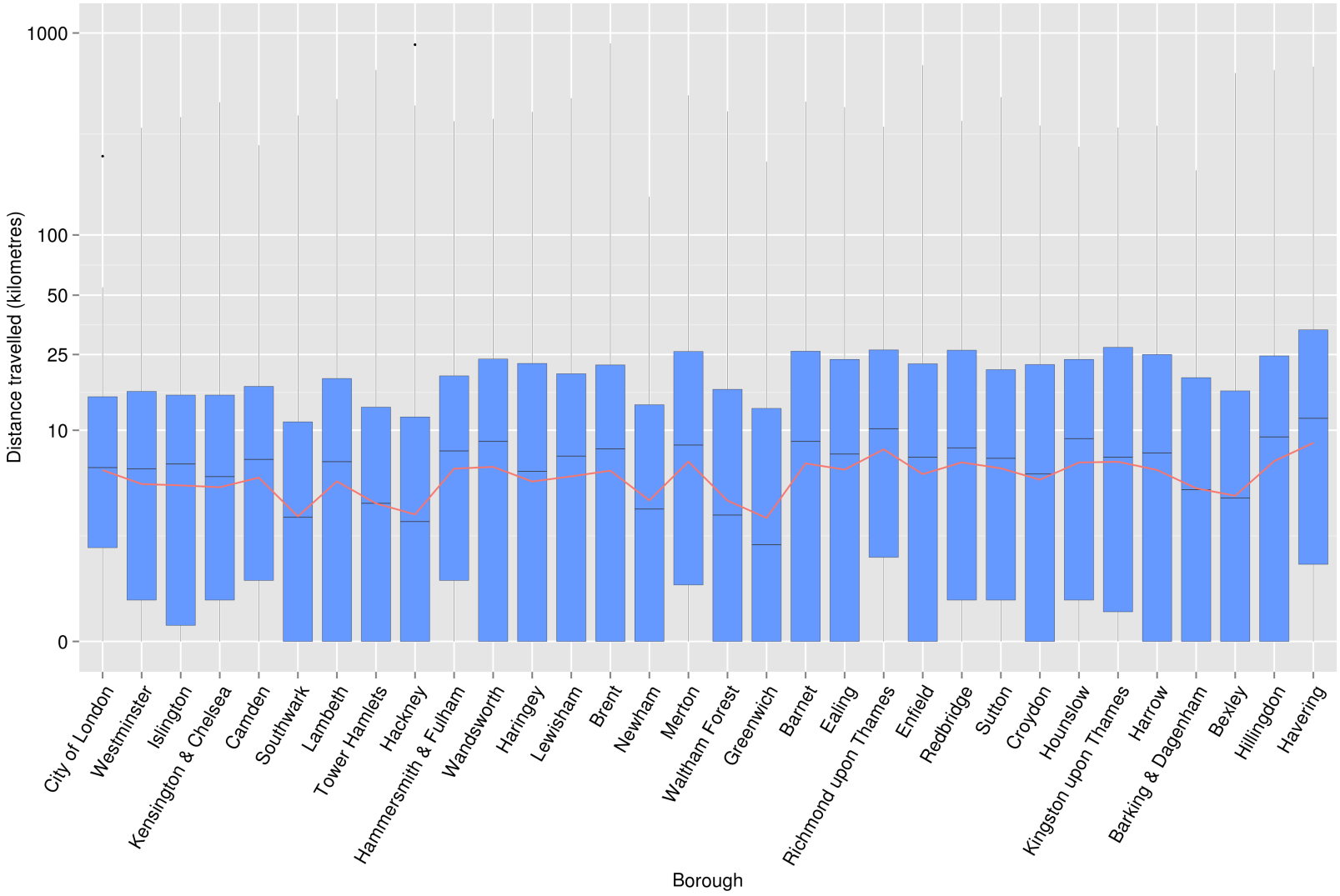


Figure 3.16: Mean distances travelled by Borough of residence, ordered from closest to centre of London (left) to furthest (right)

There did not appear to be any clear pattern between where people live and the distance that they travel each day. As the Boroughs were plotted in order, if it was true that people living in outer London Boroughs travel more, then the medians and means of the graph should rise from left (closest to centre of London) to right (furthest from centre of London). The Borough with the lowest mean was actually found to be Greenwich, perhaps reflecting that the people who live in Greenwich tend to work more locally (this could be investigated using the dataset). The boroughs with the highest mean travelling distance were Havering, Richmond and Merton.

3.5.7 Transport mode choice by age group

Focusing on transport mode choice, the percentage of time that each person spends within each transport micro-environment (bus, car, cycling, train, underground and walking) was calculated (Figure 3.17).

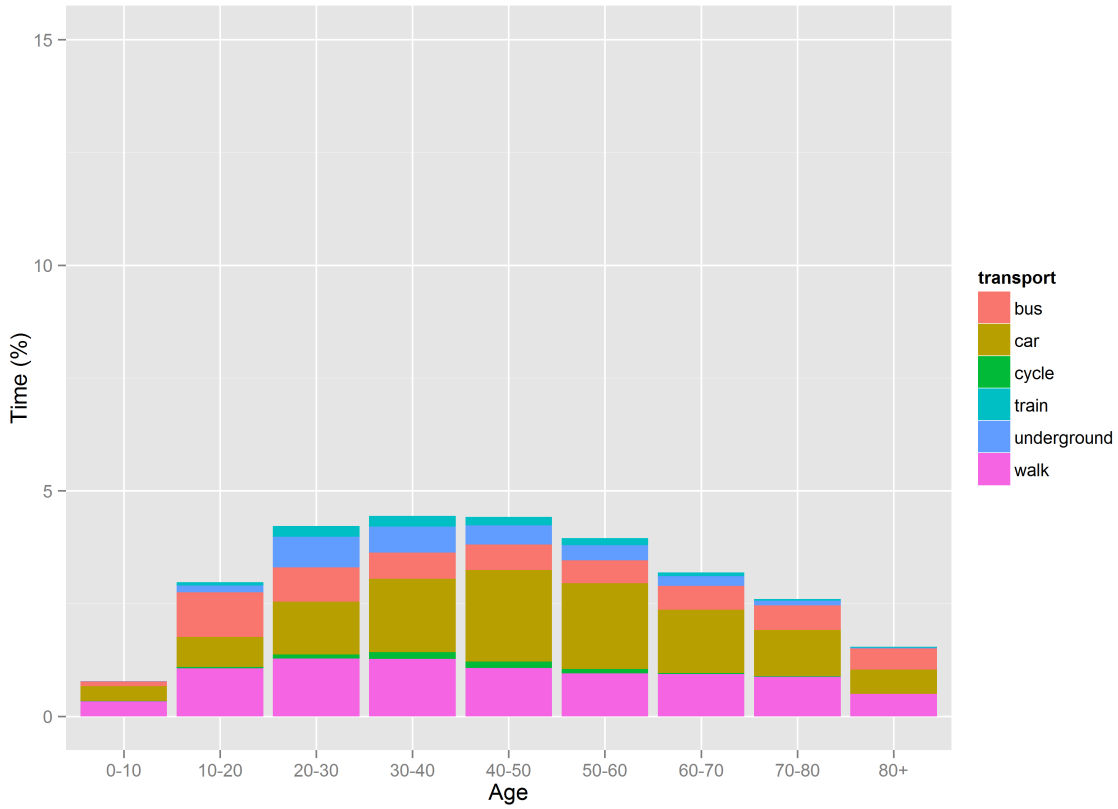


Figure 3.17: Percent of typical day using transport modes by income bracket

This graph agrees well with Figure 3.14 (travel distances by age group) which showed a bell-curve like distribution, with a leaning towards the upper age groups. Additionally however

Figure 3.17 shows how use of transport modes changes with age. Except for the 0–10 age group, bus transport is used the same amount by all ages. Walking has a similar pattern, being used frequently by people even up to the ages of 80 and above. Cycling is popular with the 20 to 60 year old's, but then hardly used by people older than that. The underground results are interesting in that it is almost never used by anyone in the 0-10 age bracket, or the 80+ age bracket, despite free travel being granted to people in those age groups.

3.5.8 Time near residence

One of the main motivations behind this research was to examine where people spend time, and therefore accumulate their exposure to poor air quality; and thus whether using address or postcodes (or dynamic methods) for exposure assessment was valid. To examine this a 1km buffer was created around each residents recorded residential address (Shown in Figure 3.18), then the percent of time that each occupant of each house spends within that buffer was calculated (that time being the start of each subjects day, then end, and any time between trips which ended/started in that area). The results are shown in Figure 3.19.

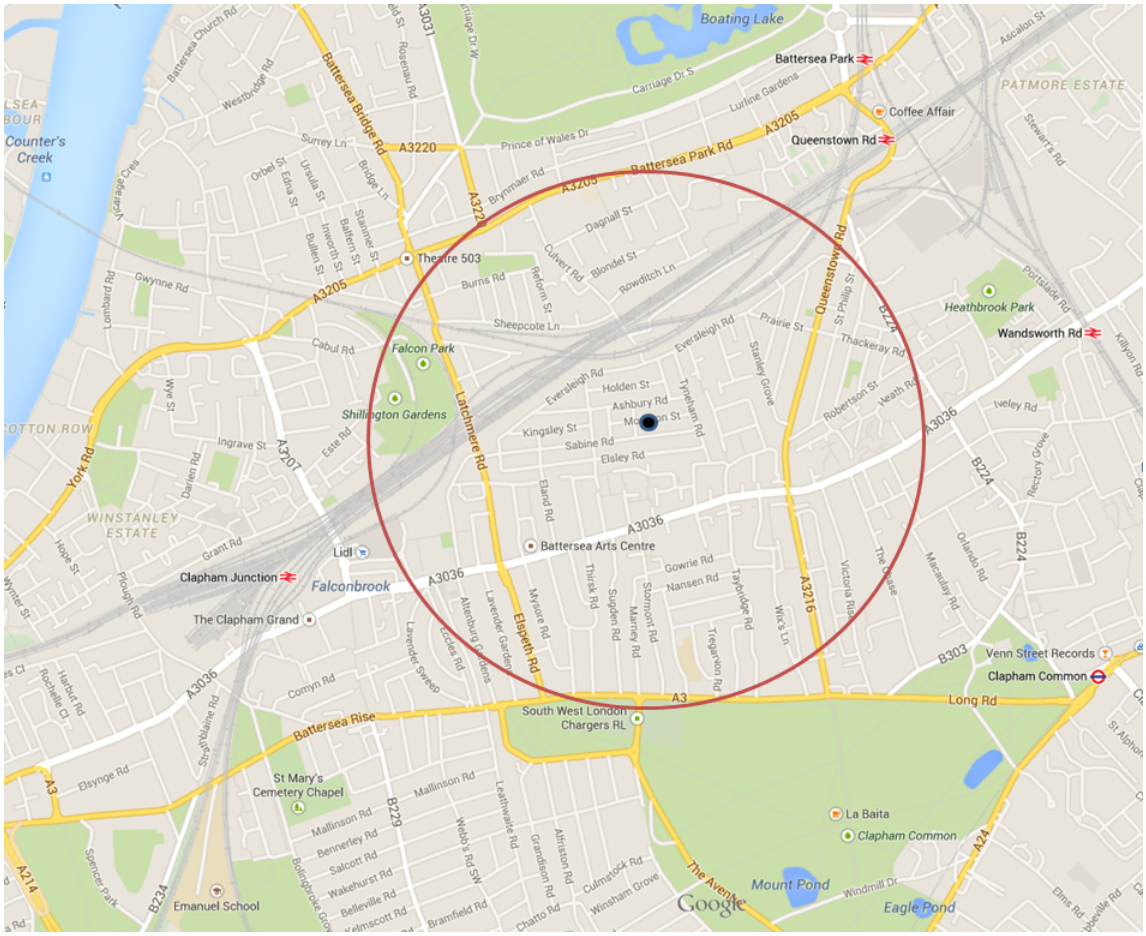


Figure 3.18: Example of a 1 km buffer around residential address

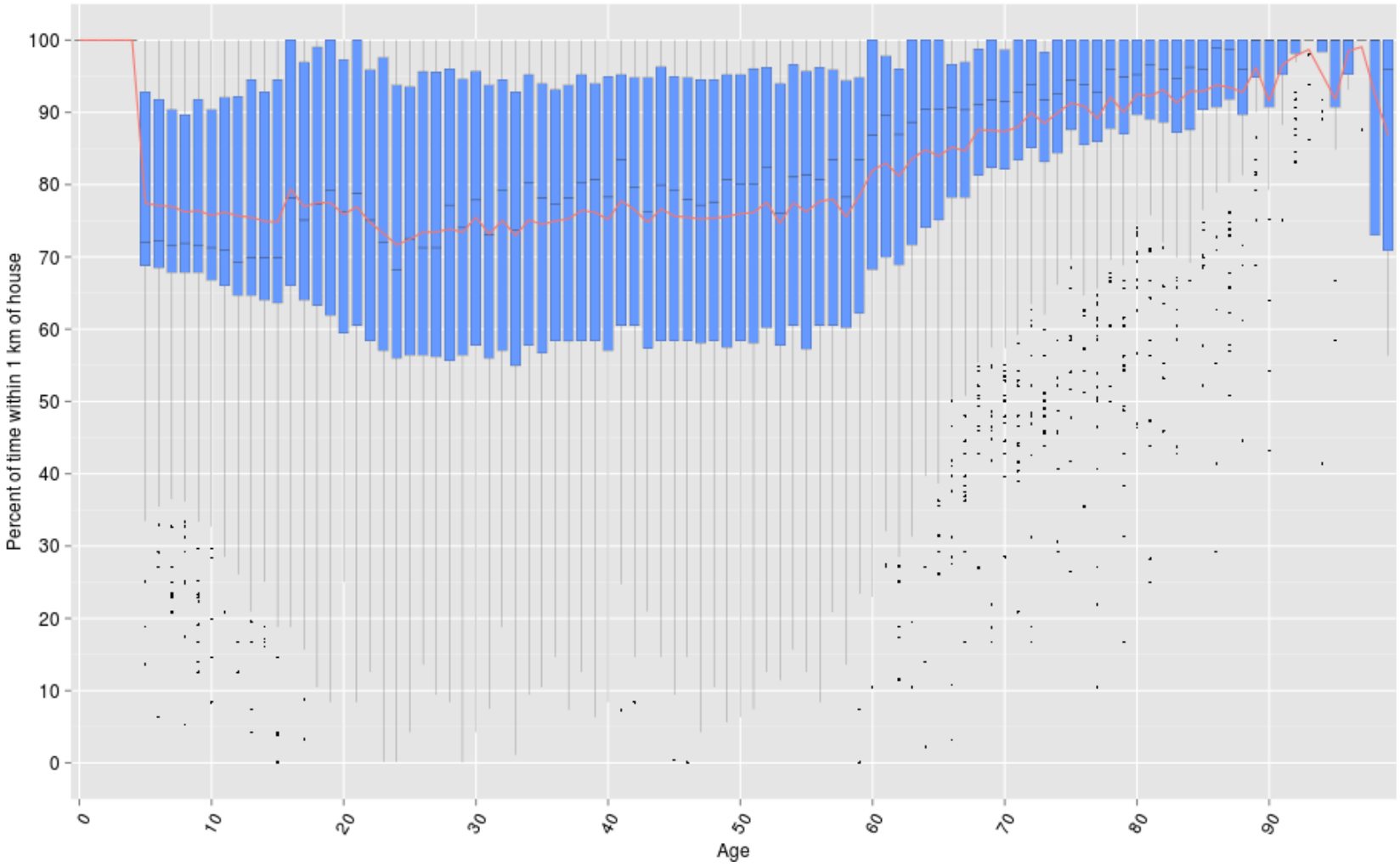


Figure 3.19: Boxplot of percent of time within 1 km of home address, means shown by red line

Figure 3.19 shows an interesting pattern of time spent within subjects local neighbourhoods varying by age group. Considering the mean line (red), time spent at home hovers between the 70% and 80% mark with small variations between the ages of 5 and 60. From age 60 onwards time spent within the local neighbourhood steadily increases to 90% by age 70, and mid 90s from there onwards (with a sudden dip in the 98 and 99 year olds, however this seems likely to be due to small numbers of subjects of this age who happened to be quite active on the day of the survey).

For illustrative purposes three simple animations of the data–set were made. The first two using the QGIS TimeManager plug-in, and the third using javascript. They can be viewed online at the following URLs:

- A webpage with an interactive slider, whereby the user can view the time–activity points of the trips the subjects of the LTDS make, on a minute by minute basis, with the points colour-coded according to the mode at that point–time:
  - `http://www.londonair.org.uk/research/dynamic_london/mode_animation.html`
- An animated GIF file of the movements of the LTDS, clipped to show specifically London
  - `http://www.londonair.org.uk/research/ltds_uk/traffic_london.gif`
- An animated GIF file of the movements of the LTDS, showing the whole of the UK
  - `http://www.londonair.org.uk/research/ltds_uk/traffic_uk.gif`

These illustrations allow a much better understanding of the data–set than was possible before. Instead of rows of an database, the data can be viewed and understood. When used in presentations to colleagues and external stakeholders (such as the TfL staff who collect the LTDS data), it was particularly useful in helping them become more familiar with the results of the process. Particularly interesting to note in the animations is the spatial distribution of the LTDS subjects, with their movements on the day of the survey clearly not limited to the geographical area of London. Also the number of people who travel from London to outside of London for work or otherwise.

3.6 Discussion

Processing and manipulating the LTDS with the methods and tools discussed in this chapter proved complicated, particularly in writing the SQL and R scripts to interrogate the APIs, and then the spatial and temporal interpolation techniques to re-define the data in the

resolution required. However having initially attempted to build custom transport networks using PostGIS and PgRouting (the PostgreSQL extensions), the methods attempted were much more productive in that they allowed a far higher percentage of routes to be solved, and with much higher accuracy (visual comparison).

The dataset that has been produced is, to my knowledge, unparalleled in this area of study. Few studies have produced similar datasets, a notable one being Dhondt et al. (2012) and their 5 million subjects in the 'FEATHERS' dataset, however the temporal resolution of that model was only hourly and the spatial resolution down to 16 km by 16 km zones of Flanders (Belgium) rather than exact coordinates. Conversely, it could be argued that the hybrid_location table actually be more highly resolved that it is. Time intervals of one-minute were chosen for convenience, when 30 seconds or even 1 second could have been used. However this seemed the most sensible choice on the basis that the dataset is already very large. Further research to assess the impact of perhaps changing to 30 second intervals could be worthwhile. The linked demographic data to the people and households of this data was also very useful in examining travelling patterns and mobility by different characteristics, which has been done in many studies before however normally by using crude measures such as euclidean distances between households and recorded workplaces, rather than by using routing solutions.

The limitations of this approach include the lack of finely-detailed control over the routing process and solution. Whilst most of the APIs allowed for different transport modes to be selected, and some of them give options such as 'avoid major roads', they do not allow cost-factors to be dynamically assigned to roads or for the network to be manipulated such as removing specific routes from the choices available. Whilst not an issue at this stage of this research, a situation can be envisaged whereby low exposure routing would be desirable, and this will prove difficult. Similarly, there is currently no way of validating the routes that were chosen by the subjects, as being representative of the actual routes taken. For some transport modes this seems less of a worry than for others. A subject taking a tube from London Bridge to Elephant and Castle for example has very little choice of their route (without significantly adding time to their journey), and much would be the same for other journeys using public transport. Private modes such as driving and walking would be less certain, a driver taking a detour due to roadworks for instance, which the routing API (being ran 2 years later) would not be aware of.

The way that time inbetween trips is considered might be leading to an over-estimation of the amount of time that subjects spend indoors each day. This is because when subjects finish trips this method presumes that they are then inside until the next trip begins. Time spent not travelling but in outdoor spaces, is therefore neglected. Intersecting an Ordnance

Survey land-use map with the dataset could lead to improvements in this area of the model. Further efforts could also be made to 'recover' discarded data which was removed in the data-cleaning process, for example subjects with miss-aligned trip start and end times were removed, but further investigation may have been able to resolve these issues.

Going forward, this dataset will now be referred to for simplicity as the LTDS-Expanded dataset, or the LTDS-X for short.

3.7 Conclusions

The aim of this research was to process and characterise the LTDS to create a new data–set that can be used as an input for a hybrid exposure model. Spatially-enabled databases with custom scripts were wrote to do this, R was used as an interface with various APIs, and the 'hybrid_location' table was created. Extensive quality checking and data cleaning was undertaken. The final dataset allows interrogation of the daily activities of people in London, which can be interrogated on a fine spatial and temporal detail, aswell as by various demographics. The main findings from this data–set were as follows:

- There are peaks in travel between 8am and 10am in the morning, and between 5pm and 6pm in the evening (As would be expected), however substantial travel also occurs between those hours.
- Men travel more each day than women (mean of 18.28 km, compared to women with 13.89 km).
- Households with a higher average annual income, are inhabited by people who travel more than households with a lower average annual income.
- The distances that people travel each day increases as they get older, peaking around the ages of 38-42. It then steadily declines, however much more gradually than it rose. By late 80s, people travel very little.
- There appears to be no pattern to where people live compared to the distances they travel each day
- The amount of time that people spend in transport each day peaks in the 30-50 age categories.
- Car use is the most popular form of transport amongst all age groups over 20 (until that point walking is the most popular). Walking is the second most popular, followed by the London Underground.

- People over 80 rarely use the London Underground.
- The amount of time that people spend within 1 km of their home, when analysed by age group, has a similar but slightly different pattern as the distance that people travel. Between the ages of 5 and 20, the mean is around 77%, which then drops to the low 70% mark for people aged 20 to late 50s. From 60 onwards, the time at home gradually increases up to high 80s.

This initial exploration and validation of the key characteristics of the data were important to confirm that it was suitable for use in further research. It will now be used as an input to a hybrid exposure model to estimate subjects exposure to air pollution.

4. Dynamic exposure modelling

4.1 Aim

Model exposure to $PM_{2.5}$ and NO_2 of the LTDS-X subjects, and compare with traditional exposure methods.

The following paper was published in 2016 - it is primarily based on the research conducted in the previous and current chapters.

Smith, J. D., Mitsakou, C., Kitwiroon, N., Barratt, B. M., Walton, H. A., Taylor, J. G., Beevers, S. D. (2016). The London Hybrid Exposure Model (LHEM): Improving human exposure estimates to NO_2 and $PM_{2.5}$ in an urban setting. Environmental Science & Technology, acs.est.6b01817. `https://doi.org/10.1021/acs.est.6b01817`

4.2 Objectives

- Link population movement data to air quality data
- Incorporate I/O ratios and micro-environmental factors
- Create a postcode comparison dataset
- Create a address-point comparison dataset
- Create a monitoring site comparison dataset
- Analysis

4.3 Background

Chapter One (Modelling Londoners movements) focused on processing, checking and exploring the spatial data that was created from the LTDS, leading to the LTDS-X. The daily journeys of around 45,000 people were recreated on a fine spatial and temporal scale from survey data. This data-set was created to allow investigation of exposure in urban environments, and also miss-classification, as described in Section 2.4 (Dynamic exposure & health studies), namely the differences between assigning someone a static exposure value based on their home location, postcode, nearest monitoring site or otherwise, compared with a model which considers all the micro-environments and varying concentrations during the subjects daily movements.

This chapter will link the LTDS-X dataset to the CMAQ-Urban dataset (described in brief in Section 4.3.1, 'CMAQ-Urban' below), and then undertake micro-environmental modelling for when the subjects are indoors or in transport. The completed model of exposure is henceforth referred to as the London Hybrid Exposure Model or LHEM. The LHEM will be used to explore exposure variations within the subjects, to attempt to identify and understand any patterns, and to consider exposure missclassification that may be occurring using standard exposure methods. Potential policy uses of the LHEM, and ways in which it might be useful to the general public, are then discussed.

4.3.1 CMAQ-Urban

The 'Community Multi-scale Air Quality model' (CMAQ) is an ongoing open-source project, coordinated by the US-EPA Atmospheric Science Modelling Division. It is made up of a variety of software packages and processes for simulating air quality. To quote from their website, "CMAQ combines current knowledge in atmospheric science and air quality modelling with multi-processor computing techniques in an open-source framework to deliver fast, technically sound estimates of ozone, particulates, toxics, and acid deposition" (United States Environmental Protection Agency (2014)). The three main components are as follows (CMAS Centre (2014)):

1. A meteorological modelling system for the description of atmospheric states and motions
2. Emission models for man-made and natural emissions that are injected into the atmosphere
3. A chemistry-transport modelling system for simulation of the chemical transformation

and fate

The Atmospheric Dispersion Modelling System (Urban) (ADMS-Urban), is a separate air quality model which is distinctive as it is able to model a range of scales, from street to city, taking into account emissions sources such as traffic, industry and domestic sources. It incorporates advanced algorithms for the height-dependence of wind-speed, turbulence and stability to produce predictions. It also includes information on street canyons and mixing introduced by road traffic re-suspension (Cambridge Environmental Research Counsultants (CERC) (2014)).

KCL-Urban (Beevers et al. (2013)), developed in the ERG of KCL, outputs annual mean air quality predictions of CO, NO_2, O_3, PM_{10} and $PM_{2.5}$ on a regular 20 m x 20 m grid. This is combined with a CMAQ regional scale model to create 'CMAQ-Urban' providing predictions of the same pollutants at the same scale on an hourly temporal resolution. The two models are equipped with similar capabilities in that KCL-Urban is quick to run and can provide details of the poor air quality sources at any location within it's domain. By combining these two (CMAQ and KCL-Urban) the result is a model which is partly deterministic (uses fundamental physics and chemistry), but provides the same spatial detail as KCL-Urban. Importantly, it is therefore capable of predicting hourly concentrations (Beevers et al. (2013)). An example of the output of CMAQ-Urban is shown in Figures 4.1 and 4.2.

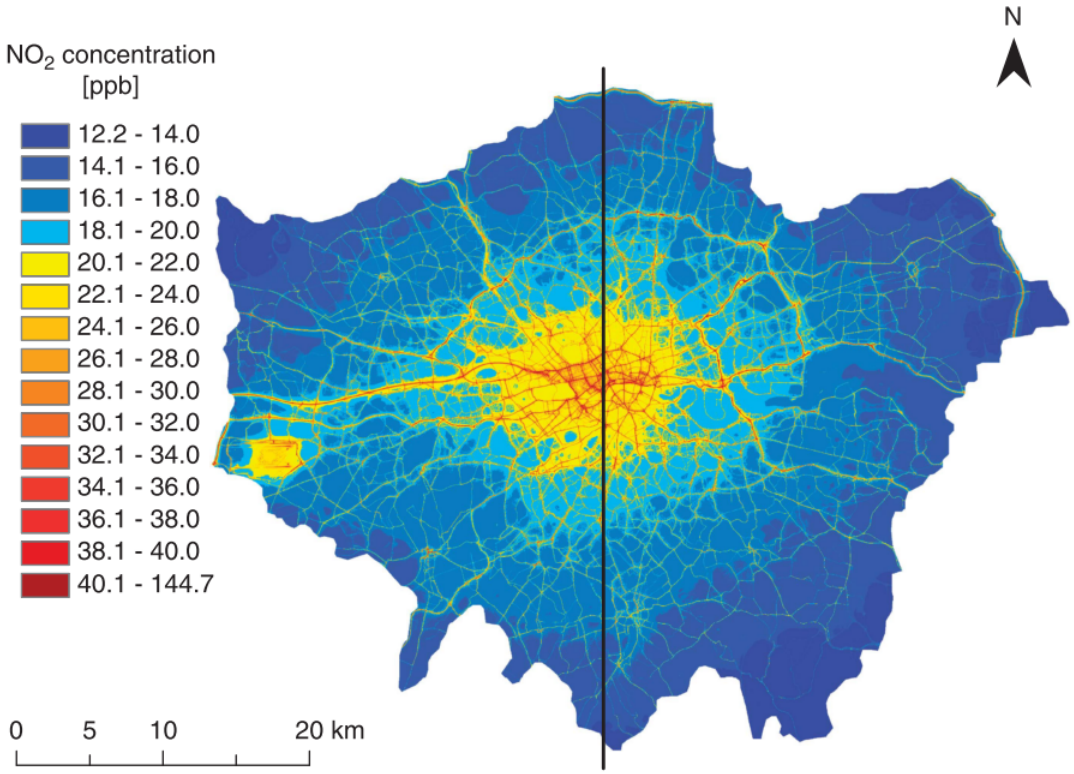


Figure 4.1: Annual mean NO_2 concentrations in London for the year 2008 predicted onto a regular grid of 20 m x 20 m using the KCLurban model.

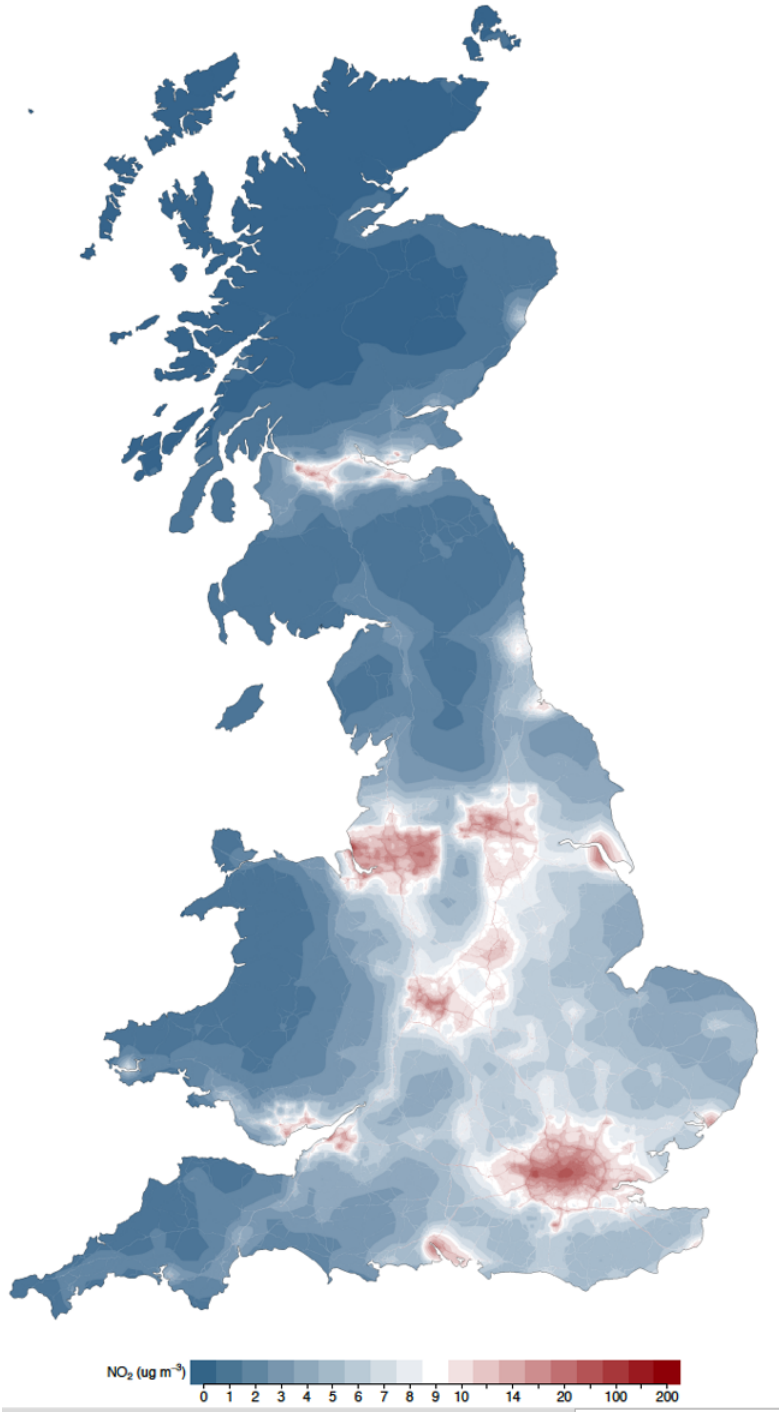


Figure 4.2: Annual mean NO_2 concentrations over England, Scotland and Wales at 20 m x 20 m resolution from CMAQ-Urban

CMAQ-urban was submitted to the UK Model Intercomparison exercise run by the UK Government department DEFRA (Carslaw et al. (2013)), where it performed well against other urban and regional models with r values of 0.9 for NO_X and NO_2, and 0.77 for $PM_{2.5}$.

4.4 Methods

In order to calculate the 'static' exposure estimates which will be compared to the LHEM estimates, a number of input data-sets and methods/processes needed to be completed. First the methods of completing the LHEM are explained, then of creating a postcode-level exposure model, followed by address-point exposure model, and finally monitoring-site exposure model.

4.4.1 Running the London Hybrid Exposure Model

4.4.1.1 Linking ltdsx to outdoor concentrations

At this start of this chapter the CMAQ-Urban air quality model was introduced. It was explained how this model outputs daily (weekday/Saturday/Sunday), hourly concentrations for a 20 metre by 20 metre grid covering the UK. The concentration of a range of pollutants at the location of each individuals minute-by-minute location over 24 hours was therefore now extracted from this model output i.e. the LTDS-X was linked to a CMAQ-Urban layer by time and location. Due to the memory and processing power needed to run the CMAQ-Urban model, and the language that it is written in (Fortran with SQL inputs), linking CMAQ-Urban directly to the PostgreSQL database that the LTDS-X data is held in was not possible so a CSV file of the LTDS-X in lat/long format along with a timestamp was exported from the database. To avoid duplication at this stage, a SQL query was written to only export unique points. To explain, if someone stayed in the same location between 07:00 and 07:45 on a Saturday, only one point was exported as the temporal resolution of the CMAQ-Urban model is monthly, daily, hourly and thus the concentration at that point would not change during that time-frame. If however the same person was in constant movement between 07:00 and 07:45, then 45 points were outputted (as concentrations would be different in different locations for each minute). By taking the temporal resolution of the CMAQ-Urban model into account when doing the export query, the number of points that needed concentrations extracting from the CMAQ-Urban model was reduced from around 64 million (45,079 people multiplied by 24 hours multiplied by 60 minutes) to just over 4 million. Dr Kitwiroon of the Environmental Research Group at King's College London processed this dataset with the CMAQ-Urban model, and returned a CSV file with additional columns for a range of pollutants (including crucially NO_2 and $PM_{2.5}$). This CSV was then re-imported back to the PostgreSQL database and linked to the LTDS-X data.

4.4.1.2 Modelling for in-building exposure

While LTDS-X subjects are indoors, the air quality that they are exposed to is different to outdoor air. A method to estimate exposure to outdoor air when indoors was therefore required, particularly given that subjects spend so much of their time indoors. An indoor/outdoor (I/O) model to estimate concentrations inside buildings, by taking ratios and applying them to the outdoor CMAQ-Urban concentrations, was used to achieve this. The model that was chosen for this was developed by Dr J. Taylor of UCL, in which he assumes 15 building types derived from the English Housing Survey, and then creates building physic models using the location of the dwelling, window opening and closing behaviour, occupant behaviour, deposition rates and penetration factors. This model was chosen as it was specifically developed for London (and therefore the building archetypes are representative), was recent (2014), and due to our close relationship with Dr Taylor meaning that we were able to ask for minor customisations to the model to be completed, and for a number of repeat model runs to be undertaken to examine sensitivity of the model (not covered here). The methods are more fully described in Taylor et al. (2014), and the data was provided by personal communication with Dr Taylor (with minor customisations as mentioned to provide hourly results and the addition of NO_2 instead of just PM). The process of incorporating this dataset with our model is now briefly described.

1. Importing I/O ratios The data-set was provided by Dr Taylor as a CSV file with hourly I/O averages for each district level postcode in London. The data was re-organised slightly, before being imported to the PostgreSQL database.

2. Linking postcodes boundaries to Indoor/outdoor ratios In order to link the provided I/O ratios to the locations in the LTDS-X, the areas that each postcode covers was required (the file provided by Taylor contained a ratio, and a postcode but no geographical data). The geographical postcode data-set that was used is described in Section 4.4.2.1 (Importing postcode boundaries), and was linked to the ratios dataset using the postcode field common in both files. Figure 4.3 (below) shows the data at this step in the process for one hour of the day (with the London Underground superimposed to aid orientation of the map).

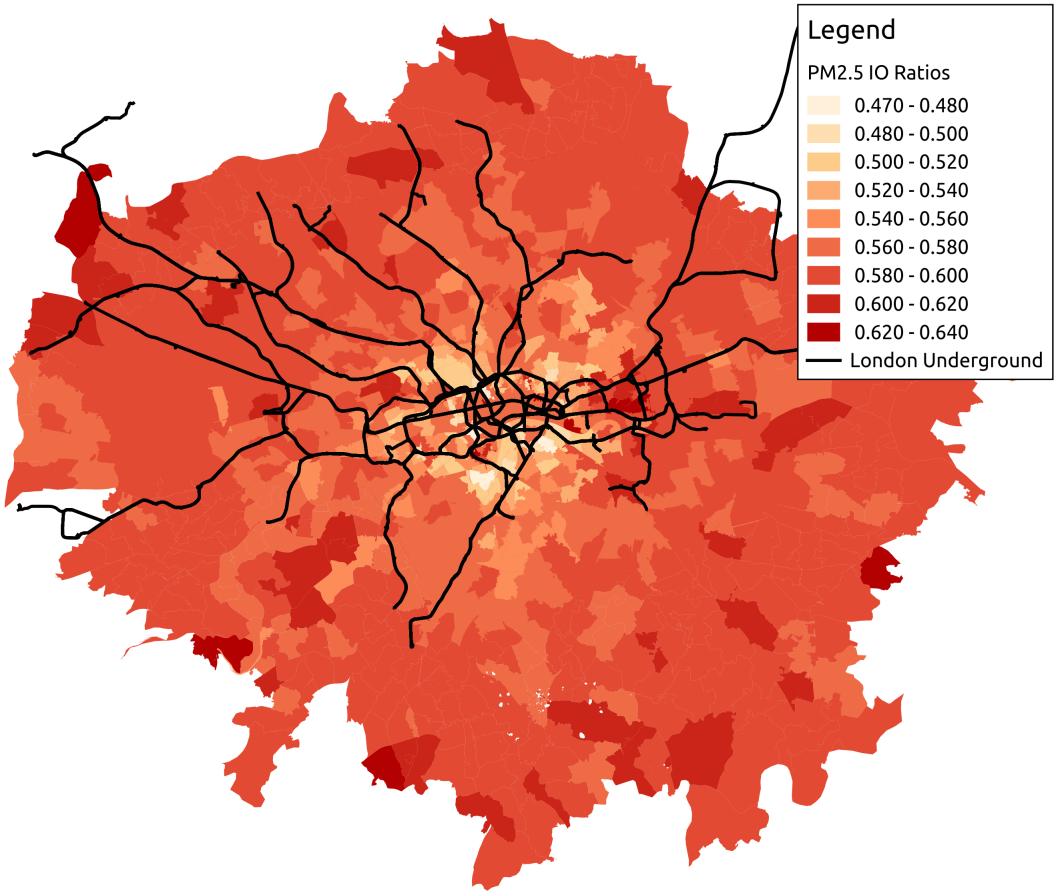


Figure 4.3: Map of average indoor/outdoor (I/O) ratios used in the LHEM. Superimposed on the map is the London Underground network to aid orientation

3. Take each point and link to the correct I/O on time/place To then link the LTDS-X data to the postcode I/O ratios, a spatial SQL query to join the two tables was written. The join was performed using the location of the LTDS-X point, as well as the time of day.

4. Multiply I/O ratio by CMAQ-Urban concentration The appropriate I/O ratio was then multiplied by the CMAQ pollutant concentration for that location, day and hour, the result being the indoor exposure at that minute. This process was repeated for all the LTDS-X data-points that are recorded as being indoors in the dataset (approximately 62 million points)

4.4.1.3 Modelling for in-vehicle exposure

Locations in the LTDS-X recorded as travelling inside vehicles needed further micro-environmental modelling to take into account that the air quality is different from the outside air. This

concept was introduced in Section 2.1.6.2 (In-vehicles) and then exposure studies that considered this area of modelling and exposure were examined in Section 2.4.3 (Transport). To calculate in-vehicle exposure in this model, the pollutant concentration (C_{in}) was derived by solving the following mass balance equation below (Equation 4.1).

$$\frac{dC_{in}}{dt} = \lambda_{win}(C_{out} - C_{in}) - n\lambda_{HVAC}C_{in} - V_g(A^*/V)C_{in} + Q/V \tag{4.1}$$

- C_{out} is the outdoor CMAQ-Urban concentration linked in Section 4.4.1.1
- λ_{win} is the air exchange rate from the windows
- λ_{HVAC} is the air exchange rate from the mechanical ventilation system
- *n* is the filter removal efficiency taking values between 0 and 1
- V_g is the deposition velocity in m/h^{-1}
- A^* is the internal surface area available for deposition
- V the volume of the vehicle
- Q is the in-vehicle particle emission rate in ug/h^{-1} (defined as the product of the re-suspension rate and the number of active passengers)

This was solved analytically and the general solution is shown below in Equation 4.2.

$$C_{in} = (C_{in_0} - \frac{b_0}{a_0}) \cdot exp(-a \cdot t) + \frac{b}{a} \tag{4.2}$$

The parameters for this model change as the subjects move, and for different vehicle types. For an example of how this module of the LHEM operates, please find a standalone piece of R code in Appendix section A.1.

For subjects locations in the LTDS-X that are described as being on the London Underground, fixed concentrations were used due to a lack of the data required to create a model of adequate spatial and temporal resolution. These fixed concentrations were derived from measurements conducted at underground platforms and on trains, in a separate as yet unpublished study within the Environmental Research Group at King's College London (Dr. Barratt, personal communication). For $PM_{2.5}$ the values of 94 μg m^{-3} (winter) and 68 μg m^{-3} (summer) were used, and for NO_x the value of 51 μg m^{-3} was used. n.b Chapter 5 aims to refine the estimation of exposure while on the London Underground.

4.4.1.4 Summary of ltdsx to lhem

The LTDS-X data is processed, as described above, using the appropriate method for each micro-environment. Once complete a dataset of 1,440 records (24 hours x 60 minutes) including time, location and exposure is output for each individual and then the 1-minute resolution data are averaged into hour of the day or full 24 hour period to obtain the typical exposure for each individual. These are then grouped or disaggregated as appropriate depending on the analysis required. Methods of exposure using 'standard' methods are now explained, for comparison with this dynamic method.

4.4.2 Creation of a postcode comparison dataset

4.4.2.1 Importing postcode boundaries

To be able to calculate annual average pollutant concentrations for each London postcode, a dataset of postcodes was required, and specifically one which contains the geographical information describing the boundary of the postcode polygon. In the UK the Royal Mail is the organisation with authority for maintaining a list of all postcodes, specifically the dataset is called the Postcode Address File or PAF. However this dataset does not contain the geographical information to link the postcodes to any other spatial data. Therefore the Ordnance Survey dataset, Code-Point Open (Ordnance Survey (2015a)), was considered. This is a dataset maintained by the Ordnance Survey, derived from the Postcode Address File, which adds the Easting and Northing of the postcode centroid to each of the Royal Mail postcodes. Although geographical coordinates now allow plotting of this dataset as points, we needed areas (polygons), and therefore this dataset was also not suitable. Fortunately the organisation Edina, part of the "EDINA and Data Library" division of the Information Services Department at the University of Edinburgh, and funded by the UK Higher Education Authorities Joint Information and Systems Committee, has derived postcode polygon boundaries from this dataset and provides the result freely to other UK Higher Education Institutions as a file called 'Code-Point with Polygons' (Ordnance Survey (2015b)). This dataset was downloaded as an ESRI shapefile and then the shp2pgsql tool used to load the shapefiles into the PostgreSQL/PostGIS database. The area of Waterloo, London is shown in Figure 4.4 to illustrate the detail of the final postcodes dataset.

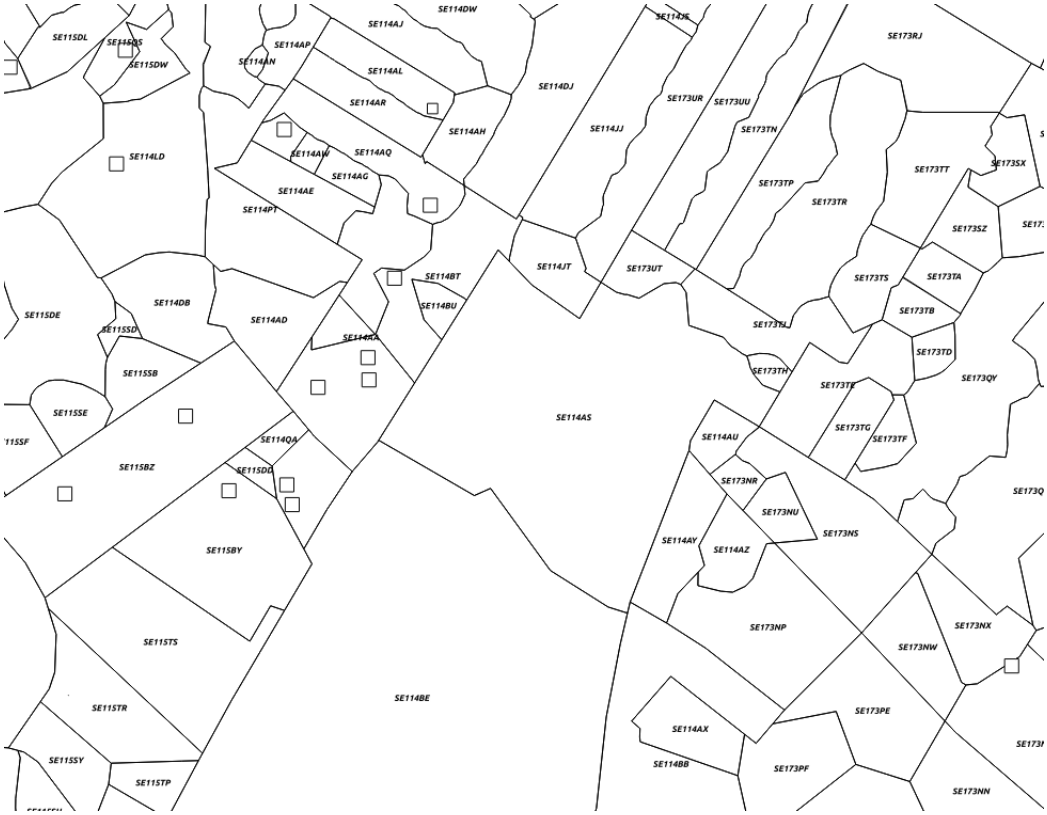


Figure 4.4: Postcode polygons from Edina Digimap

4.4.2.2 Importing CMAQ-urban annual average points

To calculate the mean annual pollutant concentration within each postcode polygon, a CMAQ-Urban output file containing annual average 2011 concentrations covering a 20m x 20m grid of London was generated by Dr Kitwiroon of the Environmental Research Group at King's College London. This was imported into the PostgreSQL/PostGIS database using the raster2pgsql tool. To demonstrate this data, 4.5 and 4.6 were produced by loading the data into QGIS and a colour gradient applied (and are shown below).

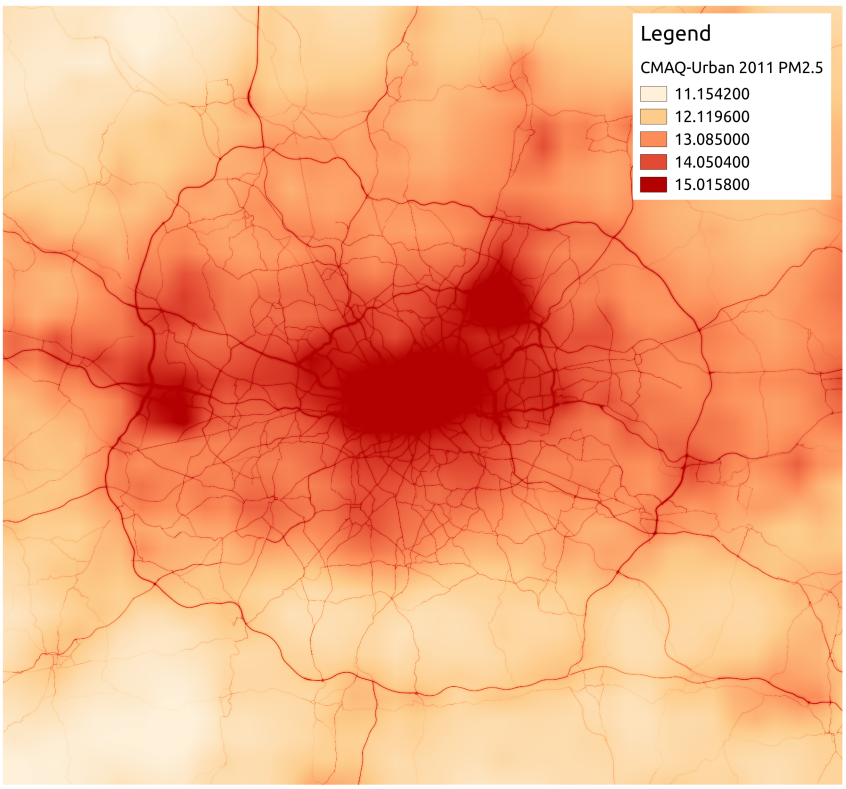


Figure 4.5: CMAQ-Urban annual mean concentration raster (2011)

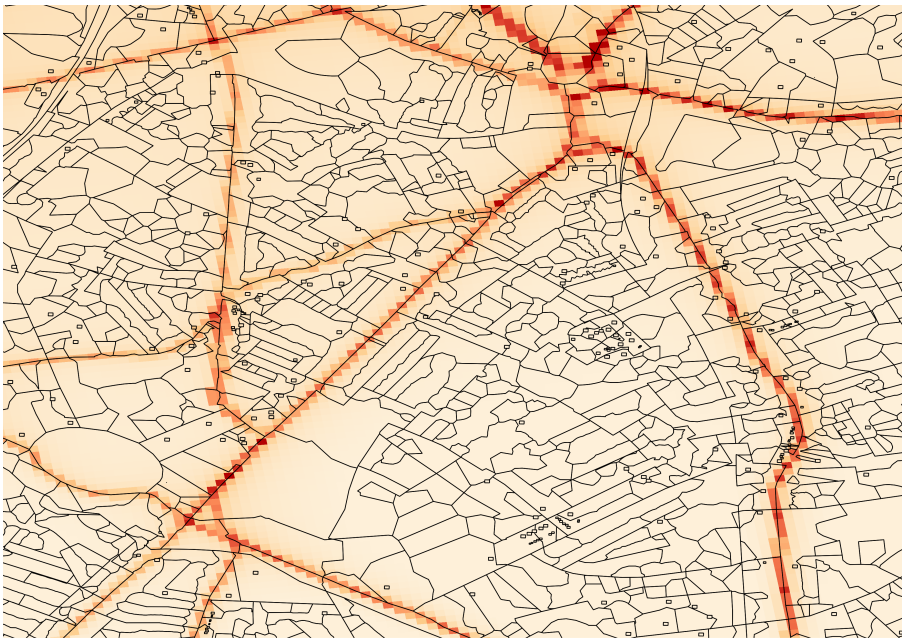


Figure 4.6: CMAQ-Urban annual mean concentration raster (2011) with postcode layer

4.4.2.3 Calculating the mean concentration for each postcode

To calculate the mean concentration for each postcode, the mean area-weighted concentration of all the 20m x 20m cells that intersected each postcode polygon were calculated. Each of the LTDS subjects' home address (based on easting and northing) was then spatially joined with the postcode dataset to establish which postcode they lived in, and the annual mean concentration for their postcode.

4.4.3 Creation of address-point comparison dataset

To calculate the address-point comparison dataset, the home address location (Easting/Northing) of each LTDS participant was first taken, as well as the day of the week that the participant was surveyed (translated into weekday/Saturday/Sunday as this was the temporal resolution of CMAQ-Urban). This data was then extracted as a CSV and processed by Dr Kitwiroon who returned a CSV with additional columns for a range of pollutants at each location. The 24 hour average for each of the LTDS subjects addresses was then calculated for each pollutant.

4.4.4 Creating of monitoring sites comparison dataset

4.4.4.1 Monitoring site data

As discussed in Section 2.3.2 (Monitoring stations), air quality monitoring stations have often been used as a measure of exposure for a population, particularly in time-series studies (Atkinson et al. (2010)), but also in some cohort studies (Dockery et al. (1993)). Comparisons between the annual average of a London 'roadside' monitoring station and a London 'background' monitoring station were therefore chosen to compare with the LHEM results. The data for these sites was downloaded using the R OpenAir package from the London Air Quality Network (King's College London (2013)) for the sites of Marylebone Road (roadside) and North Kensington (background) (shown in Figure 4.7).

Figure 4.7: The monitoring stations and surrounding areas (North Kensington left, Marylebone Road right

Specifically, the hourly means for 2011, for each site, for $PM_{2.5}$ and NO_2 were downloaded. After checking the quality of the data (as per DEFRA guidance (DEFRA (2009)), the mean of all the hours was taken to calculate the annual average pollutant concentration at that location. The results are shown in Table 4.1:

Table 4.1: Annual mean pollutant concentrations for North Kensington and Marylebone Road monitoring sites

Site	**$PM_{2.5}$ (μg m^{-3})**	**NO_2 (μg m^{-3})**
North Kensington	16.33	35.96
Marylebone Road	24.45	97.05

4.4.4.2 Methods summary

The result of these methods is a dynamic exposure model (the LHEM), and a number of comparison 'static' exposure models. In addition to comparisons with other models, the LHEM allows detailed interrogation and investigation of the exposure of ˜45,000 Londoners (and the demographic and geographic information linked to them). Calculations can be

made to answer many questions. To give an indication of it's capabilities, some examples are listed below:

- What is the average NO_2 exposure of those under 18, compared to those over 18
- What is the average $PM_{2.5}$ exposure of people of Indian ethnic origin living in Southwark
- What is the difference between exposure taken from monitoring sites, compared to exposure using the LHEM
- What percentage of Londoners daily $PM_{2.5}$ exposure comes from their morning commute
- How much less (or more?) NO_2 is someone exposed to by working at home instead of in the office
- Which Borough of London residents have the lowest exposure
- Is household income and air quality exposure related
- Do children get most of their daily exposure from within 1km of their house
- What is the difference between exposure using address-point methods, compared to exposure using the LHEM
- During which hour of the day, do people aged between 30-40 years old get most of their daily exposure

4.5 Results

The results presented here are illustrations of potential uses, focused on exploring exposure classification to $PM_{2.5}$ and NO_2.

4.5.1 The effect of microenvironments on exposure

Section 3.5 (Results) of the previous chapter calculated the amount of time that people of different age groups spend in various microenvironments during their day. Although the results varied by age group, generally the time that people spent indoors was around the 95% mark; adding weight to the sort of exposure estimates discussed in Section 2.3 (Static exposure & health studies) whereby concentrations at the subjects home or general area of residence are used to investigate the negative health effects of air quality (with the caveat that they tend to take the outdoors concentration at the area of residence, rather than attempt to model or measure the indoors concentration). Using the LHEM model, we are now able to investigate the actual exposure that occurs from each micro-environment, and examine the contribution that all microenvironments make to a subjects daily exposure. Table 4.2 summarises the time spent in each micro-environment for NO_2 and $PM_{2.5}$ by age category for the 45,709 people in the dataset, and compares these figures to the exposure they accrue in the same environment during their day (as a percent of their total exposure).

Table 4.2: Time (% of day) and exposure (μg m^{-3}) in microenvironments by age category from the LHEM model

	Age category				
	Child (5-17)	**Young adult (18-29)**	**Adult (30-59)**	**Elderly (>=60)**	**Overall**
People	10856	7474	18370	8379	45079
Percent of time in micro-environment (mean, interquartile range)					
Driving	0.77, 0-0	1.36, 0-1.32	2.24, 0-3.12	1.54, 0-2.08	1.63, 0-2.08
Indoor	97.72, 96.39-100	94.94, 92.23-98.82	94.69, 92.16-98.47	96.41, 94.86-100	95.73, 93.55-100
Walking	0.86, 0-1.53	1.66, 0-2.5	1.49, 0-2.22	1.15, 0-1.6	1.31, 0-2.01
Underground & DLR	0.06, 0-0	0.73, 0-0	0.5, 0-0	0.16, 0-0	0.38, 0-0
Bus	0.53, 0-0	0.94, 0-0.56	0.66, 0-0	0.63, 0-0	0.67, 0-0
Cycle	0.02, 0-0	0.07, 0-0	0.1, 0-0	0.01, 0-0	0.06, 0-0
Train	0.03, 0-0	0.24, 0-0	0.2, 0-0	0.06, 0-0	0.15, 0-0
Motorcycle	0, 0-0	0.02, 0-0	0.04, 0-0	0, 0-0	0.02, 0-0
Percentage of daily NO_2 exposure from micro-environment (mean, interquartile range)					
Driving	3, 0-0	5.32, 0-4.37	8.52, 0-12.62	5.64, 0-6.94	6.21, 0-7.49
Indoor	92.01, 87.35-100	82.04, 71.22-96.46	81.2, 70.58-94.99	87.97, 81.58-100	85.02, 75.21-100
Walking	2.64, 0-4.36	5.42, 0-8.41	4.66, 0-7.24	3.3, 0-4.64	4.08, 0-6.28
Underground & DLR	0.21, 0-0	2.49, 0-0	1.72, 0-0	0.57, 0-0	1.31, 0-0
Bus	1.97, 0-0	3.61, 0-1.86	2.52, 0-0	2.2, 0-0	2.52, 0-0
Cycle	0.07, 0-0	0.26, 0-0	0.39, 0-0	0.05, 0-0	0.24, 0-0
Train	0.07, 0-0	0.64, 0-0	0.54, 0-0	0.15, 0-0	0.38, 0-0
Motorcycle	0, 0-0	0.08, 0-0	0.19, 0-0	0.02, 0-0	0.10, 0-0
Percentage of daily $PM_{2.5}$ exposure from micro-environment (mean, interquartile range)					
Driving	1.3, 0-0	2.34, 0-2	3.81, 0-5.38	2.52, 0-3.15	2.77. 0-3.34
Indoor	95.89, 93.97-100	87.77, 82.96-98.04	88.59, 84.75-97.54	93.4, 91.47-100	90.98, 87.88-100
Walking	1.39, 0-2.36	2.51, 0-3.8	2.28, 0-3.4	1.74, 0-2.48	2.02, 0-3.06
Underground & DLR	0.44, 0-0	5.22, 0-0	3.55, 0-0	1.19, 0-0	2.71, 0-0
Bus	0.9, 0-0	1.56, 0-0.9	1.09, 0-0	0.99, 0-0	1.11, 0-0
Cycle	0.04, 0-0	0.12, 0-0	0.18, 0-0	0.02, 0-0	0.11, 0-0
Train	0.04, 0-0	0.35, 0-0	0.3, 0-0	0.08, 0-0	0.21, 0-0
Motorcycle	0, 0-0	0.03, 0-0	0.08, 0-0	0.01, 0-0	0.04, 0-0

The results of comparing time in microenvironments to exposure in those same microenvironments shows that the contribution of the indoor environment is very important to exposure – people spend ˜95% of their time indoors. However when taken as a percentage of their daily exposure, it's importance is diminished, people accumulated ˜85% of their daily NO_2 and ˜90% of their daily $PM_{2.5}$ while in that micro-environment.

In contrast to this, the time that people spend in transit is small, but becomes more important when daily exposure to pollutants is considered. Across the population time spent driving is less than 2%, but over 6% of NO_2 exposure, and travel on the London underground accounts for less than 0.5% of time, but contributes almost 3% of $PM_{2.5}$ exposure.

The variation of time in micro-environments and exposure in micro-environments between age groups varies. For NO_2, children and the elderly accumulate more of their daily exposure indoors (92.01% and 87.97%) than young adults and adults (82.04% and 81.2%) reflecting the differences in time they spend in that environment and the pollutant concentrations they are exposed too during their day. This pattern is similar for $PM_{2.5}$ exposure.

With regard to which transport modes contribute most to exposure, for $PM_{2.5}$ the ranking in this model is driving, underground & DLR, walking and then the bus, with all other transport modes less than 1% of daily contribution to exposure. For NO_2 the ranking is slightly different, being driving, then walking, then the bus, then the underground & DLR – reflecting the varying levels of pollutant types in different environments. Noticeably when comparing age groups, young adults get ten times more of their daily $PM_{2.5}$ exposure (5.22%) from the underground than children do (0.44%), and four times more than the elderly (1.19%). Comparisons between active and passive travel are also interesting, for example when looking at the subjects overall, passive travel constitutes 6.84% of daily $PM_{2.5}$ exposure, compared to 2.85% of their time, but for active travel these figures are 2.13% and 1.19% respectively, meaning that on a minute-by-minute basis, active travel results in lower exposure. This pattern is similar for NO_2, where 10.52% of exposure comes from passive travel in 2.85% of their time, compared to 4.32% of exposure from 1.19% of time.

4.5.2 Comparing methods of exposure estimation

As discussed extensively so far in previous chapters, the primary function of the development of the LHEM is to consider the variation and potential exposure miss-classification which occurrs in the use of different exposure metrics. Table 4.3 below summarises (mean, median and inter-quartile range) the exposures of the 45,079 people in the LTDS dataset using five different exposure metrics to provide side-by-side comparison. All except the LHEM work

on a static-outdoor basis, whereas the LHEM attempts to model movements and the effects of micro-environments.

Table 4.3: Comparing results of exposure methodologies (n=45,079, concentrations in μg m^{-3})

exposure model	mean	median	inter-quartile range
$PM_{2.5}$			
Background monitoring site	16.33	12	8 - 19
Roadside monitoring site	24.45	22	14 - 32
Postcode	13.49	13.53	13.14 - 13.84
Residential address	13.54	13.62	12.99 - 14.16
LHEM Model	8.48	8.23	7.80 - 8.66
NO_2			
Background monitoring site	35.96	30.08	19.1 - 49.18
Roadside monitoring site	97.05	90.25	61.12 - 126.5
Postcode	34.56	34.59	31.20 - 37.59
Residential address	34.34	34.45	30.65 - 38.29
LHEM Model	13	12.34	10.82 - 14.64

For both NO_2 and $PM_{2.5}$ the LHEM calculates lower exposure than any of the other exposure metrics. The roadside monitoring site method results in the highest general exposure, followed by the background monitoring site method, followed by postcode and residential address methods which are almost identical (although with residential address giving a larger inter-quartile range). Figures 4.8 and 4.9 below show histogram plots of exposure at the residential address, compared to exposure using the LHEM, for NO_2 and $PM_{2.5}$. Residential address (rather than postcode or monitoring site) was chosen for this visual comparison to the LHEM as this is currently seen as the most accurate method due to it's fine spatial detail.

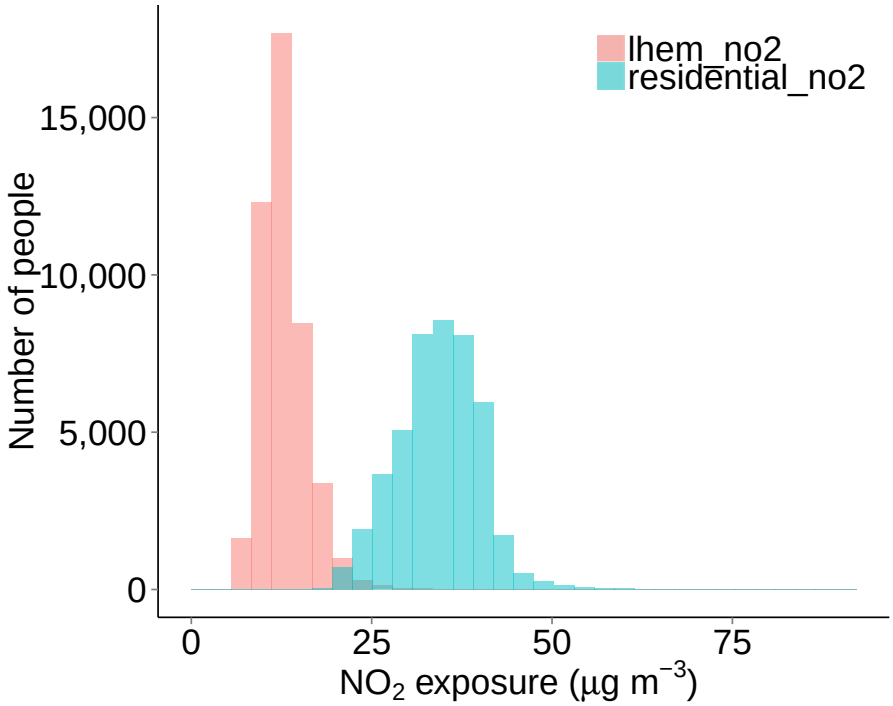


Figure 4.8: Daily mean exposure to NO_2 comparing residential address exposure with the LHEM

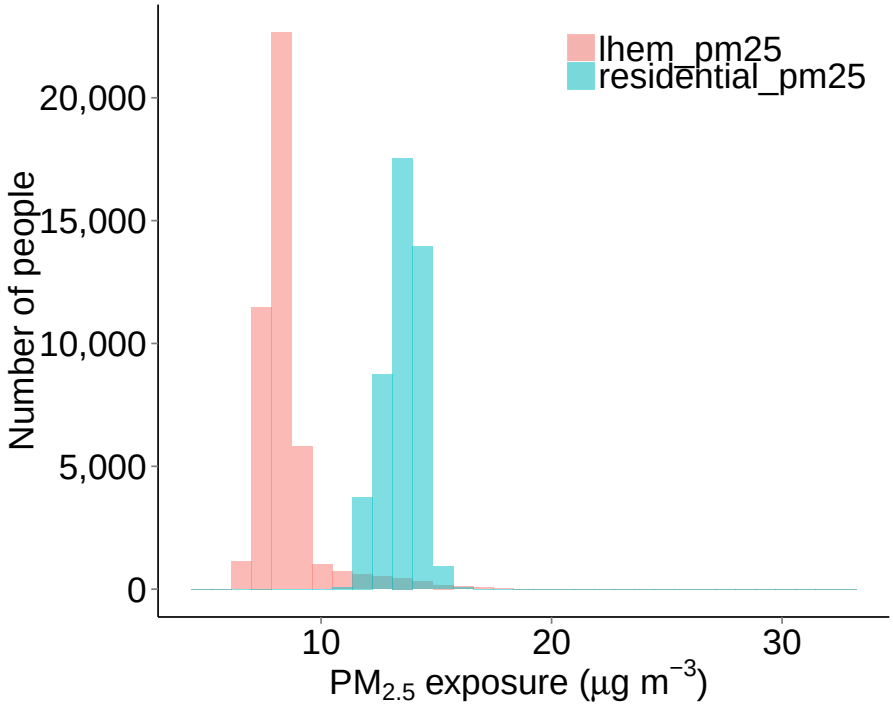


Figure 4.9: Daily mean exposure to $PM_{2.5}$ comparing residential address exposure with the LHEM

When looking at the 10th to 90th percentile range of exposures from these two methods as oppose to the inter-quartile ranges, there is little difference in the relative size of the range for $PM_{2.5}$ (2.08 (μg m^{-3}) for LHEM and 2.15 (μg m^{-3}) for residential address). However for NO_2 the range is twice as large in the residential method than it is in the LHEM (14.36 (μg m^{-3}) compared to 7.64 (μg m^{-3}))

If we now plot each individuals NO_2 and $PM_{2.5}$ exposure using the residential method, against the LHEM method, and colour-code the plots by whether the person left their house or not during the day of the survey, we can see that the change in exposure seems to be due to their movement around the urban environment away from their house (See Figure 4.10 below).

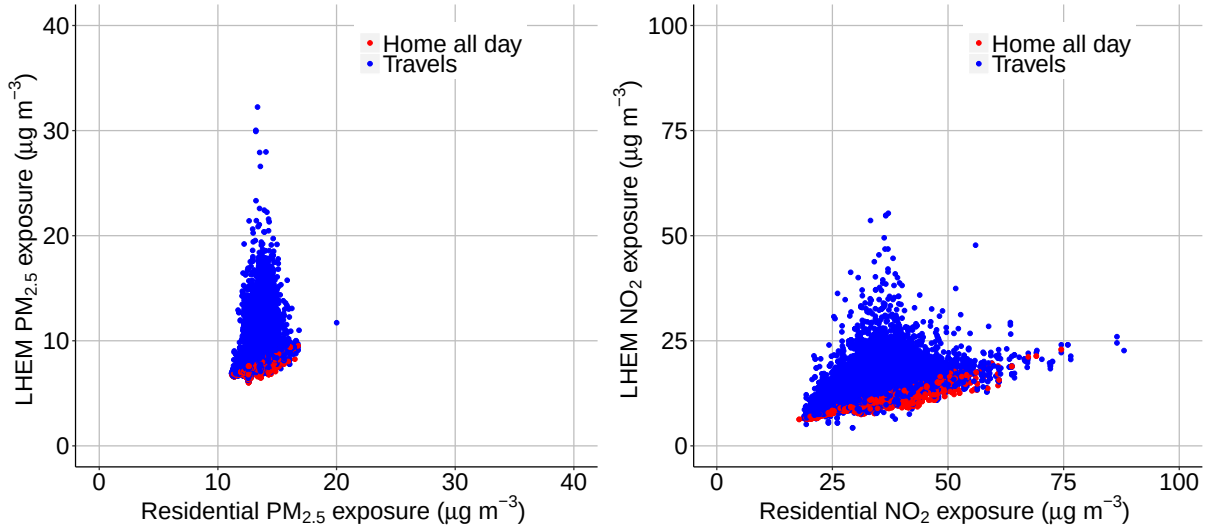


Figure 4.10: Comparing LHEM v. residential address exposure results, colour-coded by whether the subject left their house or not

4.5.3 Highly exposed people

Table 4.4 is reproduced from Section 2.1.3 (Urban environments) below for convenience and shows the acceptable mean daily and annual $PM_{2.5}$ and NO_2 concentrations as prescribed by the World Health Organisation. They are now used as a reference number with which to identify the numbers of people in the LTDS who have high average exposures when using the residential address method, and then when using the LHEM method. This comes with the caveat that the WHO limits are designed to be referenced against outdoor air quality, and so are suitable for the residential address method, but are not as suitable for the LHEM which introduces indoor and in-transport microenvironments to modelling exposure. However there are currently no regulatory limits that take account of this type of exposure modelling, and so are used as indicative values for comparison only. Specifically, the annual

average values are used as the LTDS data is designed to be typical of a days movements and activities of each person.

Table 4.4: Table of 'acceptable' WHO $PM_{2.5}$ and NO_2 levels

	Annual mean (μg m^{-3})	24 hour mean (μg m^{-3})
$PM_{2.5}$	10	25
PM_{10}	20	50
NO_2	40	200

Exposure estimates undertaken using the residential address exposure method find that 14% of the subjects have a daily NO_2 exposure higher than the WHO value of 40 μg m^{-3}. However using the LHEM model, less than 1% (18 subjects) have an exposure over this value. For $PM_{2.5}$ the residential exposure method finds that 100% of the subjects in London have an average exposure of higher than 10 μg m^{-3}., whereas the LHEM finds only 8% of subjects above this limit.

4.5.4 Exposure peaks

In the background section 2.2.2 (Long term exposure v. short term exposure) which considered the relative importance of short-term exposure compared to longer-term exposure to an individual, it was noted that hyper-short-term exposure had not been explored in epidemiological style health studies so far due to the difficulties in collecting this data and subsequent linkages to health records and outcomes. Whilst the LHEM does not have the capability to fully answer this question, it can be used to explore the variation in concentrations between micro-environments, and the time that each of the 45,079 subjects spend in environments of elevated concentrations. The table below therefore classifies the percentage of time that, on averages across the people in that age group, is spent in environments where the concentration is higher than the WHO annual mean levels for 'acceptable'. This is 10 μg m^{-3} for $PM_{2.5}$, and 40 μg m^{-3} for NO_2.

Table 4.5: Table of time of day in environments above 'acceptable' WHO $PM_{2.5}$ and NO_2 levels (I/Q range in brackets)

Age category	Percent of time in high $PM_{2.5}$	Percent of time in high NO_2
Child (5-17)	13.1% (8.8%-16.7%)	1.3% (0%-1.7%)
Young adult (18-29)	16.0% (12.5%-16.1%)	3.8% (0%-6.1%)
Adult (30-59)	15.8% (11.9%-20.3%)	3.7% (0.1%-5.8%)
Elderly (>=60)	13.7% (8.9%-17.3%)	2.0% (0%-2.6%)

In Section 4.5.2 (Comparing methods of exposure estimation) the LHEM calculated that residential address-based exposure estimates appeared to be over-estimating exposure, showing LHEM histogram plots with much lower distributions of exposure to both $PM_{2.5}$ and NO_2. However despite these lower daily averages, the LHEM also finds that, depending on age category, between 13.1% and 16% of people's time is spent in high concentrations of $PM_{2.5}$, and between 1.3% and 3.8% for NO_2.

4.5.5 Geographical missclassification

As the LTDS dataset contains the residential address of the subjects, the percentage difference between the residential and LHEM models can be calculated and then mapped to investigate whether there are any geographical patterns in the data that are not apparent from the results presented so far e.g. do people that live in North London have greater missclassification than those that live in inner London? A cumulative distribution function plot of the missclassification between the two models, expressed as a percentage, is shown in Figure 4.11 below.

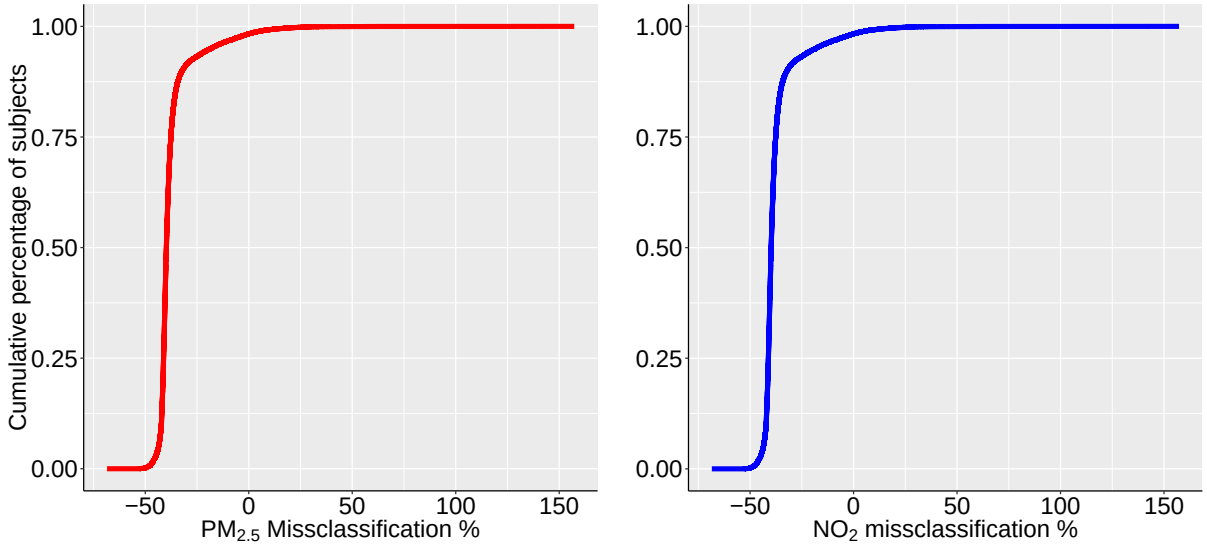
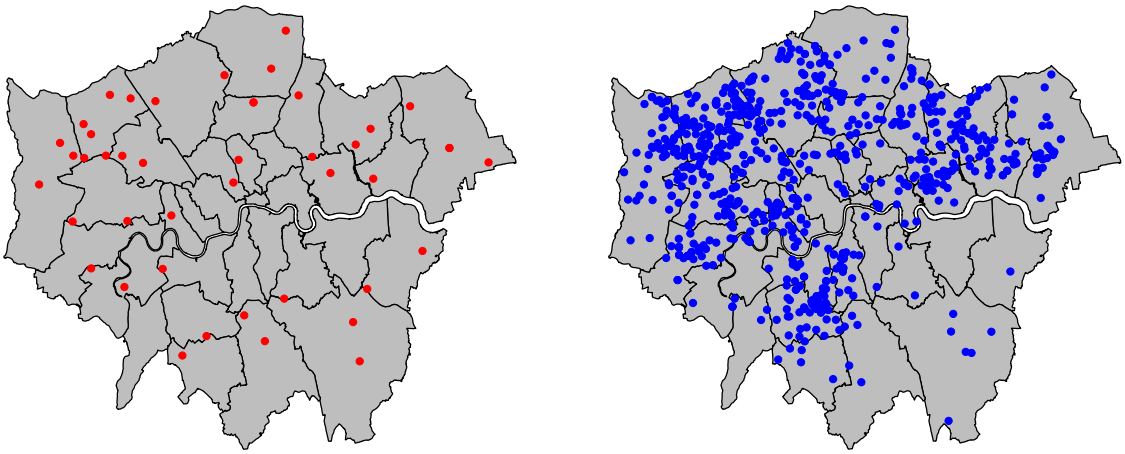


Figure 4.11: Percentage missclassification between LHEM and residential address exposure methods, plotted as a cumulative distribution plot

As is evident from these plots, for the majority of subjects in the dataset, for both $PM_{2.5}$ and NO_2, the LHEM calculates their exposure as being around 30% - 50% lower than the residential address method. There are many ways in which this aspect of the LHEM results could be interrogated, but as an example, a map of the addresses of the subjects where the LHEM calculates an **increase** in exposure compared to residential address was created to look for any obvious spatial patterns (Figure 4.12).



$PM_{2.5}$ NO_2

Figure 4.12: The residential address of subjects whose exposure increased by using the LHEM method compared to residential address

Looking at the NO_2 map first (shown right), there appears to be fewer people with an increased NO_2 LHEM exposure in the South-East of London. Whether this is a interesting result and perhaps to do with travel behaviour, or whether it is a result of the number of LTDS subjects being less in that region, is examined by plotting out all the respondents below in Figure 4.13.

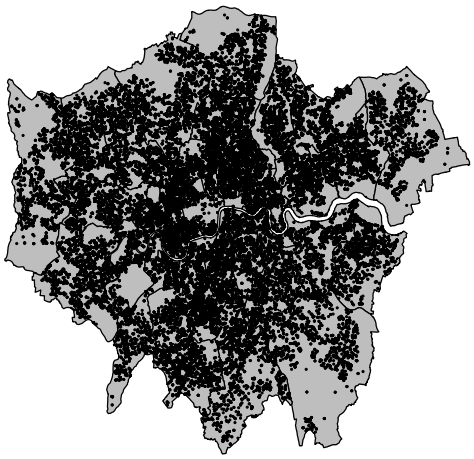


Figure 4.13: The residential address of subjects whose exposure increased by using the LHEM method compared to residential address

As can be seen, there appears to be slightly fewer people in the dataset in the South-East of London, which may explain this pattern. There could also be additional factors as work,

however this would require much more detailed analysis of this aspect of the LHEM, and some sort of regression analysis, so is not attempted in this introduction to the uses of the model.

For the $PM_{2.5}$ map, it appears that the people living in Central London are less likely to have increased $PM_{2.5}$ using the LHEM exposure method compared to the residential method, than those living further outside of central London. It seems likely that as high levels of $PM_{2.5}$ are found on the London Underground, and that people living in central London would not need to use the London Underground as frequently or for as long time periods compared to those living in outer London, that this is the reason for this geographical pattern. Although as with the NO_2 discussion above, this is not investigated further here and is merely suggested as an area for future exploration and as a demonstration of the LHEM capabilities.

4.5.6 Pollutant correlation

A further area that the LHEM may be of use in health studies is in the separation of health effects from different pollutants. Brunekreef (2007) notes that many cohort studies have looked at the negative health effects of NO_2, but questions whether NO_2 is a surrogate for other pollutants such as $PM_{2.5}$, which may actually be causing the detrimental health effects. The studies Brunekreef reviewed found it difficult to look at the relative effects of each pollutant, as they are so strongly correlated. Using our residential address exposure method we found, as other studies do, that NO_2 and $PM_{2.5}$ are well correlated with a Pearson's R of 0.90 (95% CI 0.90 to 0.90) (shown left in Figure 4.14 below). In contrast, using the LHEM (Figure 4.14, right) we see a more complicated picture of the relationship/correlation between NO_2 and $PM_{2.5}$, and a Pearson's R of 0.66 (95% CI 0.66 to 0.67). Though as the relationship is not of a linear fashion using the LHEM, Pearson's R is not a valid comparison any longer, and a more complex statistical examination would be required - probably along the lines of a regression model as briefly discussed in Section 4.5.5.

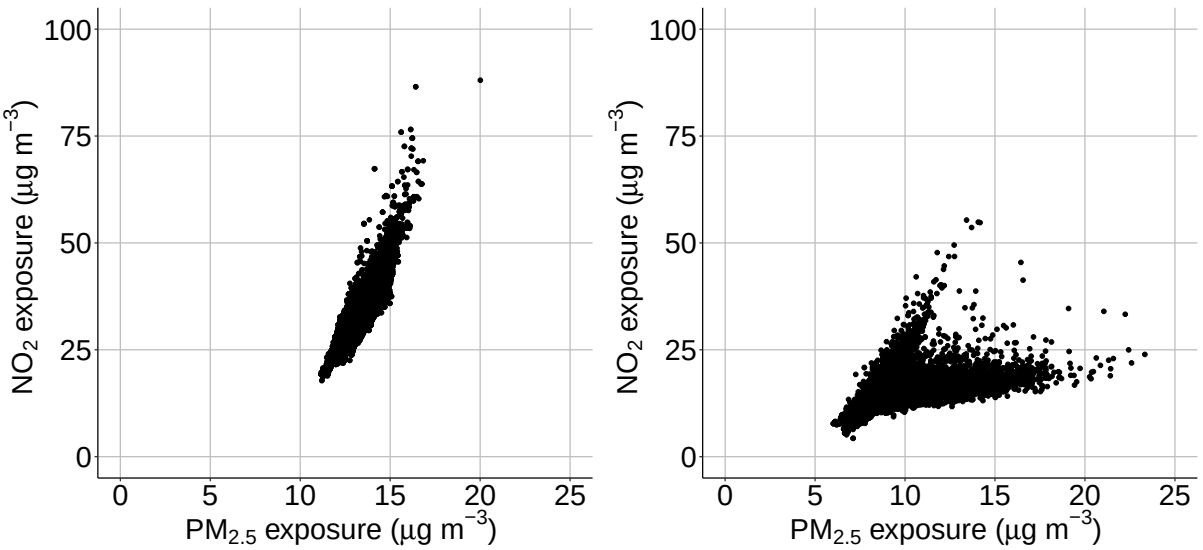


Figure 4.14: Daily mean exposure to $PM_{2.5}$ v. NO_2 using residential address exposure method (left) and the LHEM (right)

This difference between using the LHEM and residential address estimates in London is an important finding since it has the potential to separate the health effects of NO_2 and $PM_{2.5}$ as part of future public health research.

4.5.7 Susceptible groups and exposure

Studies have shown that the elderly and children are more susceptible to adverse health effects as a result of poor air quality than other age groups (Wang et al. (2015), World Health Organization (2013a)). Given this increased risk, the two tables below show summary statistics of exposure to air quality using the LHEM (Table 4.7) by age-category, for $PM_{2.5}$ and NO_2, compared to using the residential address method (Table 4.6).

Table 4.6: $PM_{2.5}$ and NO_2 residential address exposure results by age category (n=45,079, concentrations in μg m^{-3})

	Age category	**Mean**	**Median**	**I/Q Range**	**5th to 95th**
PM2.5	Child (5-17)	13.53	13.63	13.02 to 14.14	12.03-14.62
	Young Adult (18-29)	13.63	13.74	13.10 to 14.24	12.14-14.77
	Adult (30-59)	13.54	13.63	12.99 to 14.16	12.06-14.66
	Elderly (>60)	13.44	13.53	12.90 to 14.06	11.99-14.59
NO2	Child (5-17)	34.21	34.44	30.73 to 37.96	24.69-42.14
	Young Adult (18-29)	35.26	35.49	31.35 to 39.17	25.55-43.59
	Adult (30-59)	34.40	34.53	30.67 to 38.35	24.72-42.52
	Elderly (>60)	33.53	33.58	29.76 to 37.52	24.13-42.01

Table 4.7: $PM_{2.5}$ and NO_2 LHEM exposure results by age category (n=45,079, concentrations in μg m^{-3})

	Age category	**Mean**	**Median**	**I/Q Range**	**5th to 95th**
PM2.5	Child (5-17)	8.11	8.11	7.75 to 8.42	7.09-8.96
	Young Adult (18-29)	8.92	8.39	7.91 to 9.03	7.2-13.12
	Adult (30-59)	8.66	8.33	7.86 to 8.82	7.16-12.32
	Elderly (>60)	8.20	8.110	7.72 to 8.46	7.1-9.26
NO2	Child (5-17)	11.82	11.68	10.39 to 12.97	8.47-15.84
	Young Adult (18-29)	13.74	13.21	11.31 to 15.71	9.06-19.66
	Adult (30-59)	13.53	12.99	11.10 to 15.43	8.83-19.64
	Elderly (>60)	12.17	11.77	10.36 to 13.43	8.34-17.25

This next table (Table 4.8) now shows percentage change between the two exposure methods i.e. percentage change between Tables 4.7 and 4.6.

Table 4.8: $PM_{2.5}$ and NO_2 residential address exposure results by age category (n=45,079, concentrations in μg m^{-3})

	Age category	**Mean**	**Median**	**I/Q Range**	**5th to 95th**
PM2.5	Child (5-17)	-40.06%	-40.50%	-40.48% to -40.45%	41.06%-38.71%
	Young Adult (18-29)	-34.56%	-38.94%	-39.62% to -36.68%	40.69%-11.17%
	Adult (30-59)	-36.04%	-38.88%	-39.49% to -37.71%	40.63%-15.96%
	Elderly (>60)	-38.99%	-40.06%	-40.16% to -39.83%	40.78%-36.53%
NO2	Child (5-17)	-65.45%	-66.09%	-66.19% to -65.84%	65.69%-62.41%
	Young Adult (18-29)	-61.03%	-62.78%	-63.92% to -59.89%	64.54%-54.90%
	Adult (30-59)	-60.67%	-62.38%	-63.81% to -59.77%	64.28%-53.81%
	Elderly (>60)	-63.70%	-64.95%	-65.19% to -64.21%	65.44%-58.94%

Firstly, these results show that there is more variation in exposure between age groups when using the LHEM compared to using the residential address method. Secondly, they allow us to consider the different exposure of each age group from each method, and lastly they find different patterns e.g. when using the LHEM exposure method, children have the lowest exposure of all age groups to both $PM_{2.5}$ and NO_2 (and young adults the highest) – however when using the residential address method, the lowest exposure is found in the elderly instead of the children.

4.6 Discussion

The LHEM exposure model developed in this chapter, using inputs from the LTDS-X in the previous Chapter, calculates detailed spatial and temporally defined exposure estimates

on a level not seen in similar studies. The detail in the inputs separates it in particular, such as the number of subjects, the time-activity and demographic detail for them, hourly and spatially resolved CMAQ-Urban data, mass-balance modelling of some of the transport modes, and a very detailed model of indoor exposure. The closest study at this time is that of Dhondt et al. (2012), who modelled 8000 people, at 15 minute intervals, with only four microenvironments, and less detailed spatial air pollutant inputs. The results in Dhondt were quite different to the results presented here; for NO_2 they found that the static address-method was underestimating by an average of 1.2%, unlike the LHEM which finds an overestimation (13 μg m^{-3} for residential address, and 8.48 μg m^{-3} with the LHEM, meaning the address method is overestimating). Interestingly de Nazelle et al. (2013), discussed in Section 2.4.4 and in Figure 2.30, found fairly similar NO_2 results to the LHEM in their measurements-based study in Barcelona. They found that people spend 94% of their time indoors, and 6% in transport, where they accumulated 83% and 7% of their exposure. Their study population all worked at the University, and as such are likely to fall in the 18-59 age-group categories, and therefore the LHEM results fit very well with these modelled results.

As was demonstrated in the results section 4.5, the LHEM exposure model can allow detailed investigation and calculation of exposure and exposure missclassification. It was used to take a 'first glance' at the relative importance of microenvironments on exposure, the calculation of exposure for different age groups, exposure missclassification, highly exposed people, exposure peaks, whether there were any geographical patterns to exposure missclassification, correlating pollutants, and susceptible groups. Each of these areas was only briefly explored to demonstrate the potential of the model, and further analysis is needed. The model also has further uses, for example each of the LTDS subjects has many more demographic attributes that have not been explored including ethnicity, income, gender, how many cars the household owns, distance from tube stations etc. all which may influence exposure. The model can also be used for policy applications which are not explored in detail here, for example what effect would changing 5% of the subjects journeys from car to walking have on their overall exposure, and how would this vary by age group?

There are a number of ways in which the LHEM could be improved and some elements of the model that might be considered for change in the next iteration. Firstly, the number of subjects could be increased, as extra LTDS data is now available from TfL. Increasing the number of subjects is likely to increase the strength of the conclusions arrived at, however as many of the exposure results are only marginally different including new subjects may not actually change this. It is difficult to say at this stage. Another area that could be reconsidered is the use of the building I/O ratios from Taylor et al. (2014). This was an excellent dataset for this research, however it calculates average ratios at a postcode

level, for residential buildings, and therefore may not be suitable for modelling other types of buildings such as office blocks. A new iteration of this dataset might be an option, or indeed developing of an entirely new method. With regard to the transport exposure modelling, the use of set numbers for the London Underground is a drawback. The figures used for $PM_{2.5}$ were the mean of a small number of measurements on one stretch of the Underground by researchers at King's, when actually the concentrations vary on different lines, in different sections of those lines, and perhaps by time of day and season. For NO_2 the concentrations were taken from a study in Paris, and face similar problems about their applicability to London, and across the network. An intermediate step before creating a bespoke London Underground model could be be to run sensitivity analysis on the LHEM using the maximum and minimums from the measurement campaign, however refining this aspect of the model will now be the focus of the next chapter.

Perhaps most importantly for the long-term future of this type of model, and perceptions of the results that it produces, is how to validate the estimates. The most likely route to do this would seem to be using personal mobile monitoring, and then comparing this dataset to what the LHEM outputs for similar journeys and days of the subjects.

4.7 Conclusions

Results Section 4.5.1 (The effect of microenvironments on exposure) demonstrated that the LTDS subjects spend most of their time indoors, and thus understanding indoor exposure and being able to model exposure to indoor pollutants (aswell as ingress from outdoor pollutants) is important in understanding exposure in general. With the caveat that the CMAQ-Urban input to the LHEM only models exposure to outdoor pollution, it finds that people are exposed to around 85% of their daily NO_2 and 90% of their daily $PM_{2.5}$ exposure while indoors - although this varied slightly by age group, with children and the elderly accumulating more of their daily exposure from the indoors environment that adults and young adults, due to the slightly increased time that they spend indoors, and naturally a reflection of this, the increased time that adults and young adults spend in transport in environments of higher concentrations that indoors. Comparing exposure in transport modes is difficult, as except for the underground, they are thought to be a function of outdoor concentrations. However on a time/exposure basis, active travel can be seen to result in lower exposure than passive travel.

The different exposure values found in results Section 4.5.2 (Comparing methods of exposure estimation) should help epidemiologists understand that the means and ranges of exposures that are currently being used in health studies are perhaps not appropriate, and that they

may be over-estimating exposure across the population (and using incorrect ranges). The LHEM finds that estimates based on monitoring sites, postcodes or residential address are all overestimating exposure for both $PM_{2.5}$ and NO_2, by varying degrees depending on the method in comparison. Interestingly, when comparing postcode and residential address estimates across the LTDS subjects, they are found to be almost identical. Given this, perhaps the recent drive for individual residential addresses for completing exposure analysis is unneeded, and postcode estimates are sufficient to reflect the variation in exposure using static analysis methods (although studies discussed in the Introduction typically took annual averages, rather than the hourly variation that CMAQ-Urban models). When the LHEM exposure estimates are plotted against residential address estimates, with the points coloured by whether people left their house during the day, this factor appears to be the main reason for the differences between the LHEM and residential address method exposure estimates (instead of perhaps the indoor-outdoor ratios, which would only result in lower estimates but with the same overall profile). Using the LHEM method of modelling exposure, peaks, troughs and ranges in a persons daily exposure are captured which are not seen using other methods.

In results section 4.5.3 (Highly exposed people), it was demonstrated how the LHEM can investigate the numbers of people who are accumulating daily mean exposures above the WHO limits for $PM_{2.5}$ and NO_2. As static models of exposure do not take account of the subjects movements, or normally the effects of exposure being different when people are indoors, if a subjects house happens to be in an area of very high concentrations then that persons mean exposure is taken to be high. When actually the LHEM (compared to residential exposure method) demonstrates that when the subjects daily activities are taken into account, their daily mean exposure is actually (generally) lower. The result of this with the LHEM, is that much fewer subjects are found to be living in areas above the WHO limits.

As the LHEM calculates exposure on a minute-by-minute basis, Section 4.5.4 (Exposure peaks) showed how it can also be used to consider much shorter periods of high exposure in a subjects day that other models are generally unable to do. Using this advantage of the LHEM, we found that LTDS subjects are exposed to levels of 'unacceptable' (as defined by the WHO) $PM_{2.5}$ and NO_2 levels for between 13.1% and 16% and 1.3% and 3.8% of their day respectively, depending on age group.

By calculating the percentage difference between the residential address exposure method, and the LHEM exposure method, we were able to see that for most people the difference in exposure is between -50% and -20% regardless of pollutant, but how there were long-tails in the distributions and that some people were found to have higher mean exposure estimates

when using the LHEM (similar to the findings of Reis et al. (2018)). The people were plotted on a map to see if there was any specific geographical distribution to these results, but none was immediately apparent. Further spatial investigation may find associations e.g. proximity to roads or London Underground stations.

The LHEM was also able to consider $PM_{2.5}$ and NO_2 exposure correlations, used in many other health effect studies (discussed in Section 4.5.6 (Pollutant correlation). First the residential $PM_{2.5}$ and NO_2 were checked to see if the exposures in this research were similar to others, and they were found to have a R^2 of 0.90, confirming that they were. The same analysis using the LHEM resulted in a much lower correlation of 0.66. However there actually appeared to be two different correlations within the scatter-plot, and this needs further investigation. This difference between pollutants using the LHEM may allow future health studies to better be able to estimate the differences in health effects from individual pollutants.

Finally, the LHEM was used to investigate how exposure varies by age groups in Section 4.5.7 (Susceptible groups and exposure). It demonstrated how each age group has fairly similar ranges and means of exposure, and similar missclassification between the LHEM and residential exposure estimates.

5. Exposure to $PM_{2.5}$ on the London Underground

5.1 Aim

Create an exposure model for exposure to $PM_{2.5}$ on the London Underground

5.2 Objectives

- Measure $PM_{2.5}$ across the London Underground network
- Link measured air quality data to noted time-location data
- Import, process and clean other datasets for linking i.e. platform depths, station locations.
- Analysis

5.3 Background

The London Underground (otherwise known as 'The Tube') has around 402 kilometres of track covering the Greater London Area, around which 52% is overground and 48% is underground (Transport for London (2014b)). The network currently has annual passenger numbers of 1.305 billion, and is the main source of transport for the population of London.

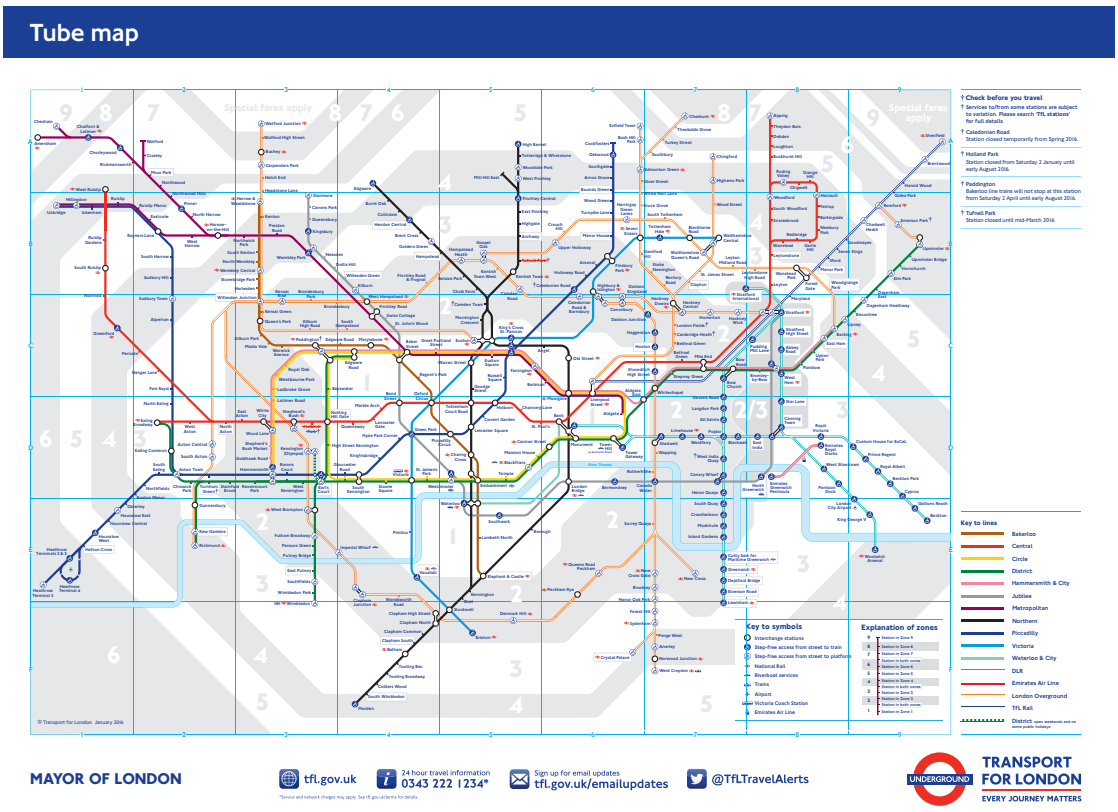


Figure 5.1: A map of the London Underground

In the previous chapter concentrations of 94 μg m^{-3} and 51 μg m^{-3} for $PM_{2.5}$ and NO_2 respectively were used as simple exposure estimates to represent exposure while the LTDS-X subjects were travelling on the London Underground network. These concentrations, particularly for $PM_{2.5}$, represented some of the highest exposures that the subjects encountered, way in excess of the concentrations found at their residential address. However the concentrations used, taken from the studies cited at the time, are the means of a wide range of measurements. For example the $PM_{2.5}$ data that was collected by Dr. Barratt (personal communication) upon which the mean of 94 μg m^{-3} was based, has variation well below and above this mean. Taking a small sub-sample of a journey between Waterloo and Bond Street on the Jubilee Line the $PM_{2.5}$ varied between 22 μg m^{-3} and 140 μg m^{-3} (1 minute averaging time).

The evidence of variation in concentrations along the lines, supporting the assertion that a more dis-aggregated method of estimating exposure on the London Underground is needed, is further strengthened by the studies discussed in Section 2.4.3 (Transport exposure) where very different concentrations were found between studies and within studies. Figure 2.27 summarised the concentrations across various studies of underground train exposure (with results for $PM_{2.5}$ on the London Underground being 202 μg m^{-3} in Adams et al. (2001a)

and 246 μg m^{-3} in Pfeifer et al. (1999)). Specifically within the Adams paper, a mean of 238.7 μg m^{-3} for $PM_{2.5}$ was found in lines below ground and a mean of 29.3 μg m^{-3} on above ground lines suggesting that whether the line is over or under ground influences $PM_{2.5}$ concentrations. Adams also noted that there was no statistical difference between concentrations at different times of day, or between the two seasons when testing occurred (summer, June 1999 and winter, February 2000).

However these studies have tended to have (i) relatively short time periods of measured concentrations (ii) study only small areas or fixed sites of the network (iii) summarise by giving an overall mean and confidence intervals, and (iv) are limited in their scope and attempts to explore the variation of concentrations e.g. by mapping their data. Additionally, the more comprehensive of these studies (in particular Hurley et al. (2003)) have framed their findings in terms of exposure on the underground network as an occupational hazard, for example comparing measured concentrations to the exposure of welders, when as we know the actual people who are being exposed to this air are from all ages, backgrounds and of varying degrees of health.

The aim of this chapter is to provide a more detailed understanding of pollutant concentrations on the London Underground. The focus will be on $PM_{2.5}$ given that their are no obvious sources of NO_2 on the London Underground and that NO_2 concentrations in the literature are found to be similar to ambient concentrations. Additionally, we did not currently have any portable measurement equipment for NO_2 available. Specifically, measurements will be taken within the tube train during journeys across the network, and then combined with a manually completed diary which will note the section of track or station that the concentrations were recorded at. As noted by Adams above, whether a train is underground or overground may be important in understanding concentration levels, so depth data will also be sourced and joined to the existing data, and then further this data will be joined to a geographic representation of the tube network. The result of this research will be a geographically defined dataset of $PM_{2.5}$ on the London Underground that can be used in modelling Londoners exposure whilst travelling on the tube.

5.4 Methods

5.4.1 Measurements

In attempting to better understand $PM_{2.5}$ levels on the tube, a TSI-Sidepak was carried while sitting in a passenger cabin and journeying around the network. One line was sampled per day (over a number of months), with the aim being to cover every section and station

of the line at least twice. For the simpler lines, with no spurs, this was a case of starting the equipment at one end, journeying to the other, changing trains, and then making a return journey. However for the more complex lines such as the DLR or the Central line where there are many different spurs and sections of lines, a pragmatic approach was taken whereby various sections were repeated to get as complete coverage as possible, resulting in some sections being repeated more than twice. Journey times are summarised in Table 5.1 below.

Table 5.1: Time spent collecting air quality measurements on the London Underground, by line

Line	**Minutes**
Victoria	97
Circle	160
Northern	246
Bakerloo	116
Jubilee	253
District	201
Piccadilly	205
Docklands Light Railway	258
Metropolitan	255
Central	222

Consideration was taken of the possible causes of variation in the concentrations, and how these might effect the results. When monitoring concentrations of $PM_{2.5}$ near roads, and to a lesser degree away from roads (background), there is normally a diurnal variation and seasonal variation that is caused by emissions from increased traffic and weather conditions respectively. This issue was taken to be of negligible importance in our sampling of the London Underground, as the concentrations seen in the pilot data and in previous studies have found no evidence of diurnal variation. Supporting this approach to the sampling is the work of Adams, also discussed in the introduction to this chapter, where they found little difference between concentrations in different seasons and times of day. Further, the effect of passenger numbers and movement on concentration levels was taken to be negligible, as particle concentration levels from this are insignificant in scale, in the same manner (Ferro et al. (2004)). Given this, re-suspension of particles from the movements of the trains seem to be the main cause of elevations in particles.

5.4.1.1 Equipment - TSI-Sidepak

A TSI-Sidepak (TSI (2015)) was used to measure $PM_{2.5}$ on the London Underground. The device uses a light-scattering technique, and is shown below in Figure 5.2. It weighs

approximately 16 ounces, and is 10.7 x 9.4 x 7.1 cm in size. Whilst in use the pump makes a low level 'hum' noise, due to the pump sucking in air. The device was placed in a backpack, and an inlet tube connected and fed out the top of the bag. For the sampling period, this backpack was then placed on the seat of a carriage (or occasionally on the researchers knees when the carriage was busy). This device was chosen for this purpose due to it's small size, ease of use, low level of noise, and use in other published personal exposure studies (Huang et al. (2015), Han et al. (2015), Yu et al. (2016)). The time-resolution for collecting data was set to one minute intervals, and at the end of each sampling period the data was extracted using the TSI software, and then loaded into a PostgreSQL database.

Figure 5.2: A TSI-Sidepak for measuring $PM_{2.5}$

According to the TSI website (TSI (2015)), the sidepak is calibrated "*to the respirable fraction of standard ISO 12103-1, A1 Test Dust (formerly Arizona Test Dust) [which] allows comparisons between measurements where the source or type of dust is predominately the same*". Therefore when this device is used it needs calibration factors calculating and applying to the data, to accurately reflect the concentrations in the environment it is recording in. Doing so is relatively simple to do when placed alongside a gravimetric measurement device, and is common in robust studies such as Torrey et al. (2015) where they found the Sidepak was overestimating by a factor of about 1.3 and Jiang et al. (2011) where the Sidepak was overestimating concentrations by a factor of 3. Within London, Dr Barratt has calculated a correction factor of 0.6 for use of the sidepak in outdoor environments (personal communication, 2016). However to our knowledge no correction factor exists for use in the London Underground, and as such this needed calculating. Briefly, as part of a separate research project at KCL, a TSI-Sidepak was placed in a small cabin on the platform of Hampstead station (a station on the Northern Line of the network). Alongside this, a ThermoFisher Partisol (ThermoFisher Scientific (2016)) was installed and both instruments had inlets pushed through holes in the ceiling of the cabin to sample the air over a three week period. The process and results are described in full in an article which will be submitted in

early 2019, but in summary the result was that a correction factor of times two should be applied when the device is sampling $PM_{2.5}$ in the tube. Or rather, this correction should be applied to the proportion of the PM that is attributable to the tube, rather than external air (which should have the aforementioned London correction factor of 0.6 applied). This process is illustrated with some simulated data in Figure 5.3 below.

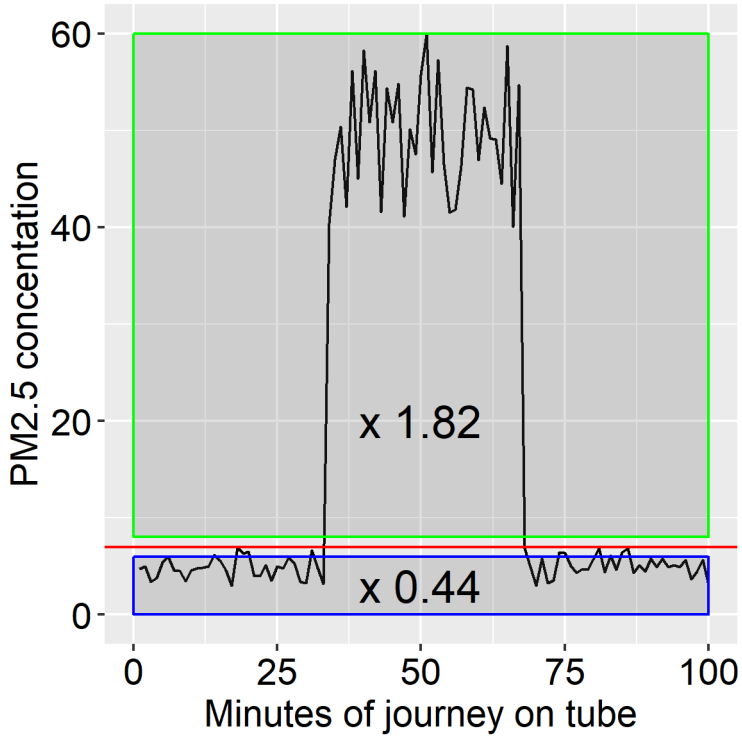


Figure 5.3: Simulated tube data showing the proportion of the data that would be scaled by 2.0 (green box) and the proportion of the data that would be scaled by 0.6 (blue box) if the London background concentration at that time was 7 μg m^{-3} (red line)

To apply this scaling factor, the daily average London background concentration of $PM_{2.5}$ was taken from the North Kensington monitoring station (part of the London Air Quality Network (LAQN) for each of the days that the sampling occurred, and this scaling method was therefore applied to all of the London Underground air quality data presented below.

5.4.2 Tube diary

In order to map $PM_{2.5}$ concentrations across the tube network, location data was needed to combine with the $PM_{2.5}$ data collected by the TSI-Sidepak. Due to around 50% of the network being underground, where GPS use is not possible, a diary was kept and then

transcribed into SQL code, before being loaded into the same database as the Sidepak data. An example of the SQL code is shown below.

```
INSERT INTO tube_diary VALUES('Bakerloo', 'Elephant & Castle', '2015-02-04 08:07:00', '2015-02-04 08:09:00', 'platform', 'floor', 1);
-- Started tube journey North
INSERT INTO tube_diary VALUES('Bakerloo', 'Elephant & Castle', '2015-02-04 08:07:00', '2015-02-04 08:07:00', 'tube', 'shelf', 2);
INSERT INTO tube_diary VALUES('Bakerloo', 'Lambeth North', '2015-02-04 08:13:00', '2015-02-04 08:13:00', 'tube', 'shelf', 3);
INSERT INTO tube_diary VALUES('Bakerloo', 'Waterloo', '2015-02-04 08:15:00', '2015-02-04 08:15:00', 'tube', 'shelf', 4);
INSERT INTO tube_diary VALUES('Bakerloo', 'Embankment', '2015-02-04 08:16:00', '2015-02-04 08:16:00', 'tube', 'shelf', 5);
INSERT INTO tube_diary VALUES('Bakerloo', 'Charing Cross', '2015-02-04 08:17:00', '2015-02-04 08:17:00', 'tube', 'shelf', 6);
INSERT INTO tube_diary VALUES('Bakerloo', 'Piccadilly Circus', '2015-02-04 08:19:00', '2015-02-04 08:19:00', 'tube', 'shelf', 7);
INSERT INTO tube_diary VALUES('Bakerloo', 'Oxford Circus', '2015-02-04 08:21:00', '2015-02-04 08:21:00', 'tube', 'shelf', 8);
```

This location data was then linked to the Sidepak data by time and date.

5.4.3 Station and line locations

To be able to investigate any spatial patterns in the data collected, we needed to add spatial attributes to the data. Whilst the tube diary and Sidepak data describe in text and numbers that, for example, the $PM_{2.5}$ levels are 37 μg m^{-3} on the stretch of line between Aldgate East and Liverpool Street at 9:32am, we do not have the geographical location of that stretch of track (or indeed the stations at either end). A geographical network of the London Underground was therefore manually created in a PostGIS database. The process is summarised below:

- Download station locations (latitude/longitude/name) from TfL as a CSV file
- Compare this list against a tube map to ensure 100% of stations were present.
- Identify missing stations (14), use Google Maps to note their lat/long, and add these to the TfL CSV
- Loaded CSV into PostgreSQL database
- Using a printed tube map, make a manual note of each section of track, including the stations that it joins, and the line of the track
- Digitise this information to a CSV (start station coordinates, end station coordinates, line ID) and load into PostGIS
- Use the PostGIS makeline() function to create a linestring feature between each station for each line as appropriate
- Visualise and error check

The key SQL code to create this is shown in Appendix section A.4.

5.4.4 London Underground station characteristics

As briefly mentioned above, the hypotheses is that line and station depths influence $PM_{2.5}$ concentrations on the network. However the data to add a 'z' attribute (depth) of each station, to test this, was not freely available via the London Datastore or similar official data sources. However a Freedom of Information request made by Hamechan Madhoo in January 2013 (`https://www.whatdotheyknow.com/request/depth_of_tube_stations_and_tube#incoming-366374`), in which TfL provided this data was found. These data were downloaded, quality checked, errors corrected, and then loaded as a CSV file into the PostGIS database, and then matched to the existing data by station name and line. Stations on the DLR are all above ground except for Bank, and were missing from this dataset, so for simplicity they were all assigned a depth z value of -10 i.e. 10 metres above ground.

5.4.5 Trains

Data regarding the types of trains that are operated on each line was also sourced from from the TfL website (Transport for London (2016)), thinking it might be useful, and is summarised in Table 5.2 below.

Table 5.2: Train type running on each tube line

Line	**Train**
Victoria	2009 stock
Circle	'S' stock (2010)
Northern	1995 stock
Bakerloo	1972 stock
Jubilee	1996 stock
District	'D' stock (1980)
Piccadilly	1973 stock
Docklands Light Railway	'B07' stock (2005)
Metropolitan	'S' stock (2010)
Central	1992 stock
Hammersmith & City	'S' stock (2010)

5.5 Results

5.5.1 Timeline concentrations

To gain an initial understanding of the concentrations and variation between each line, a time-line graph of all the data was created (Figure 5.4). The x axis was taken as minutes on the line, hence the lines that had more monitoring completed on them (sometimes due to the length of the line, sometimes due to availability of the researcher) continue further to the right of the graph than others, noting that there were repeats of each line section.

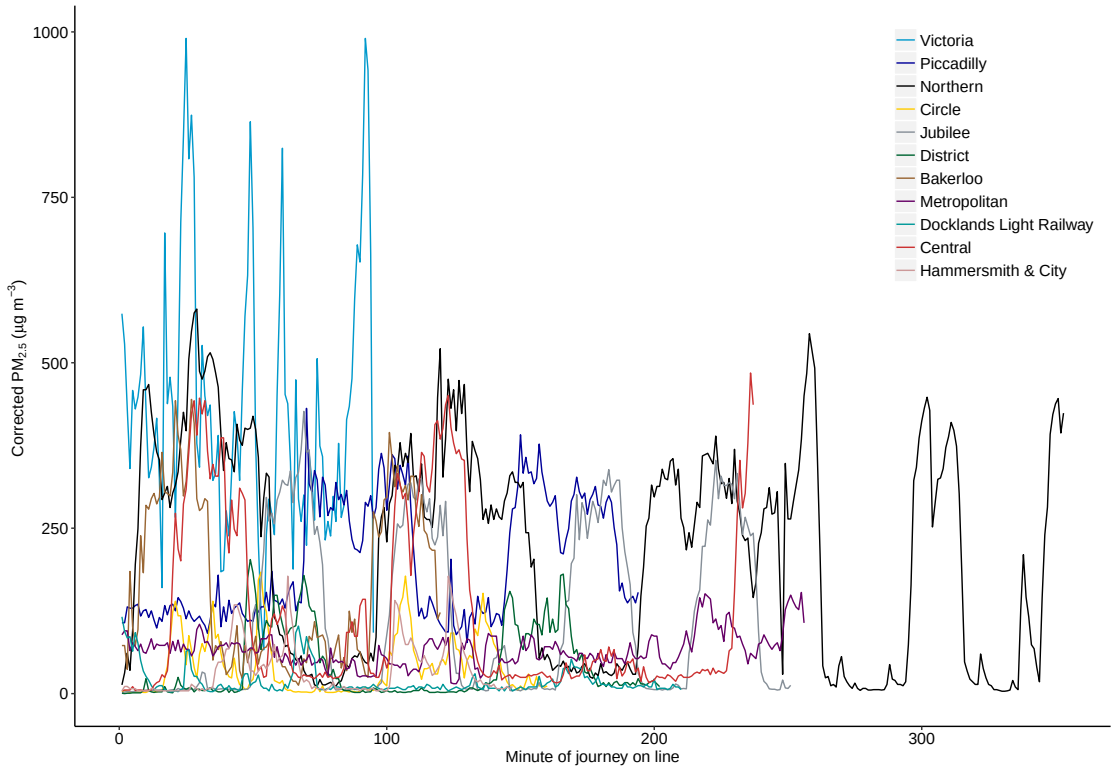


Figure 5.4: Timelines of $PM_{2.5}$ on the tube

Although plotting all the lines on top of each other in this way makes the data hard to interpret, it does immediately show the large variation in some lines, and the lack of variation in others. There are clear peaks and troughs, most apparent in the Northern line data, compared to the DLR or Metropolitan lines which although still have variation, the scale is much smaller. Figure 5.5 below plots the lines individually to enable better understanding of the patterns.

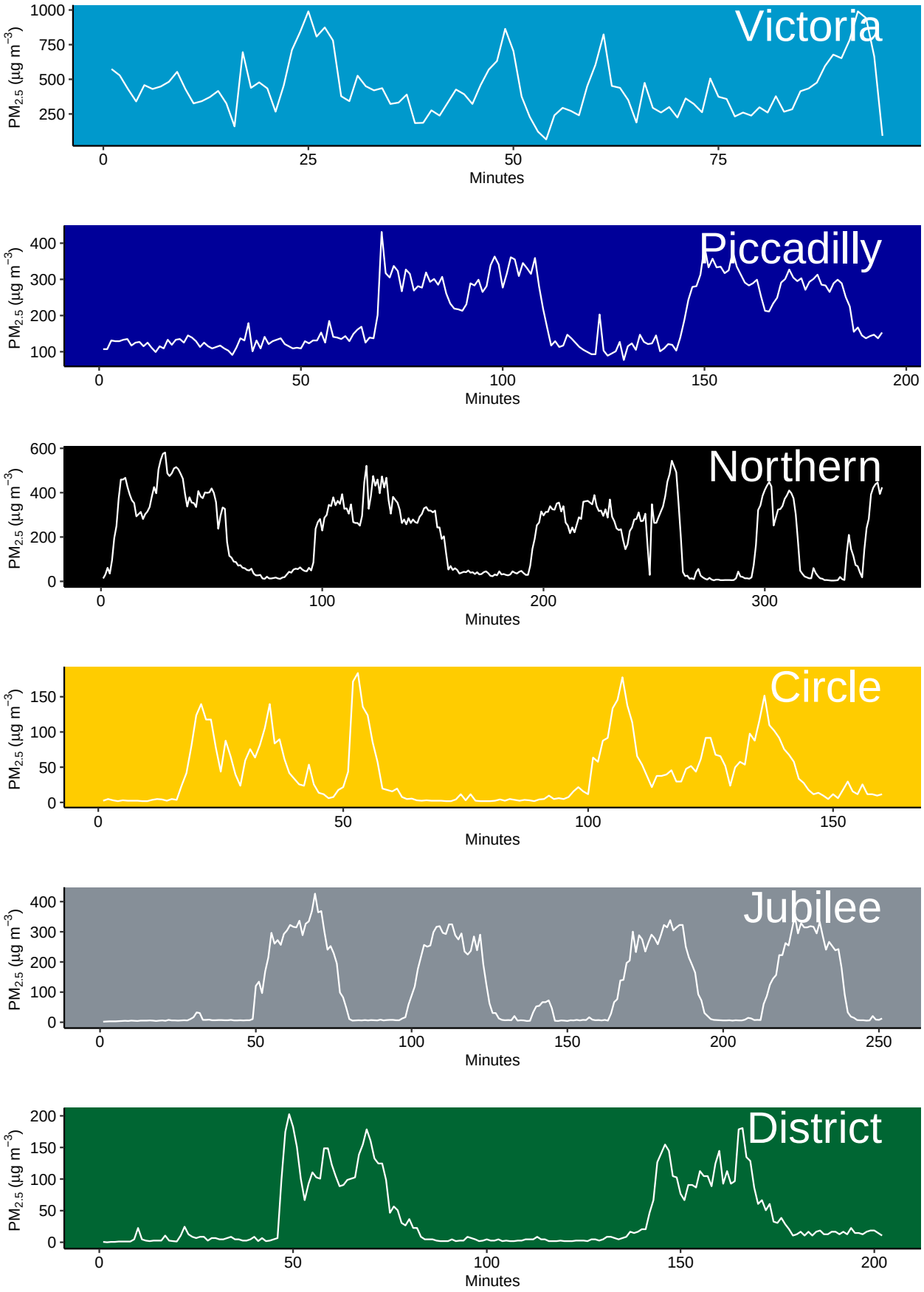

Victoria
Piccadilly
Northern
Circle
Jubilee
District
PM2.5 (µg m−3)
Minutes

Figure 5.5

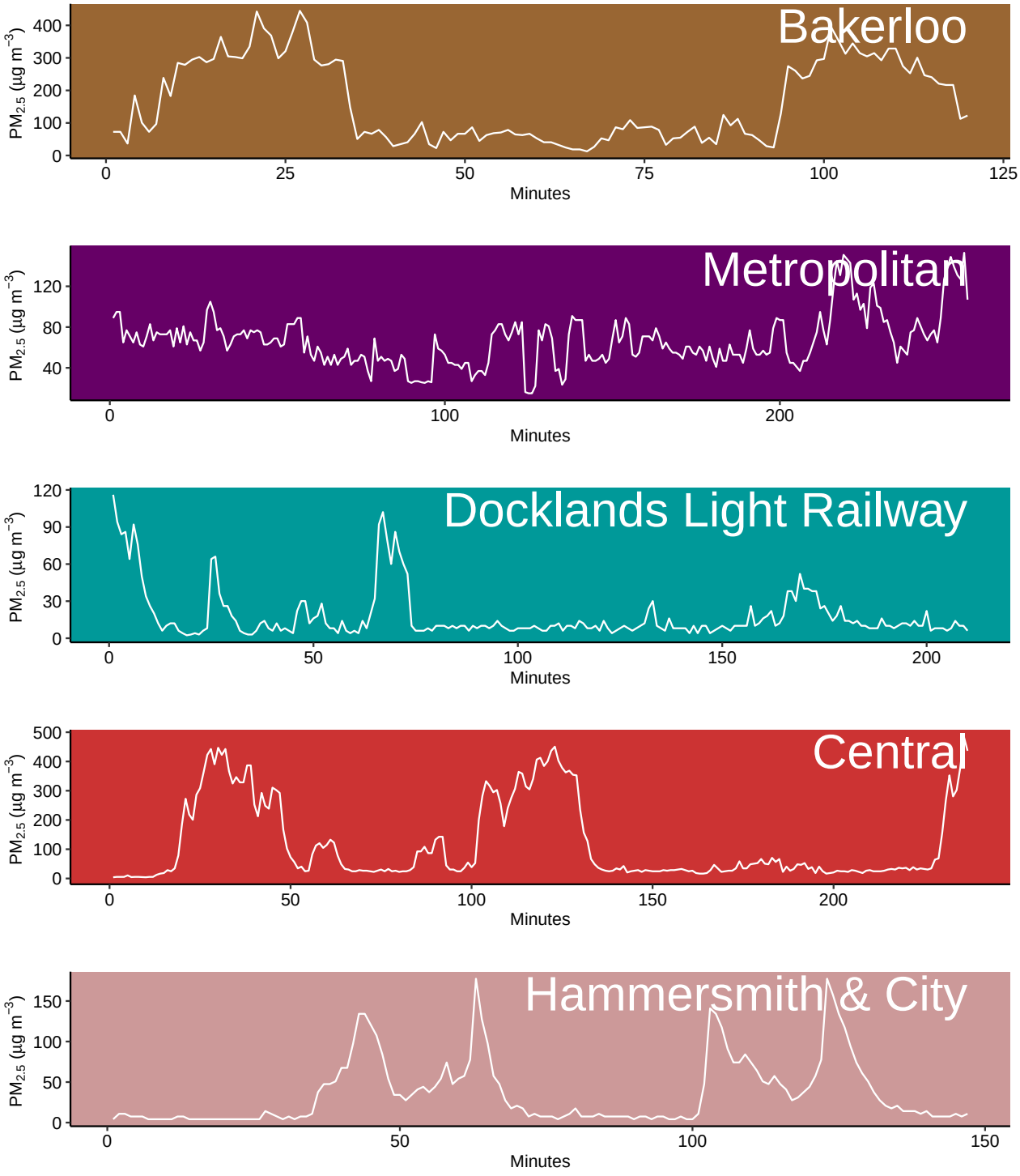


Figure 5.6: Timelines of $PM_{2.5}$ on the tube (Note differing axis scales)

We can see that the concentrations vary by line and within lines, for example the Hammersmith & City line has a maximum of around 170 μg m^{-3}, compared to lines such as the Central line with concentrations upto 500 μg m^{-3} and the Victoria line of upto 900 μg m^{-3}. Additionally, for some of the lines, there are clear patterns in the data. The Jubilee line has four clear peaks, and similarly the District line two clear peaks. Referencing these against the diary information that was collected, the peaks coincide with the train being underground, and then 'flat' low levels of $PM_{2.5}$, when the train was above ground and exposed to ambient air.

5.5.2 Line averages

Box and whisker plots were created in Figure 5.7 to compare concentrations between lines.

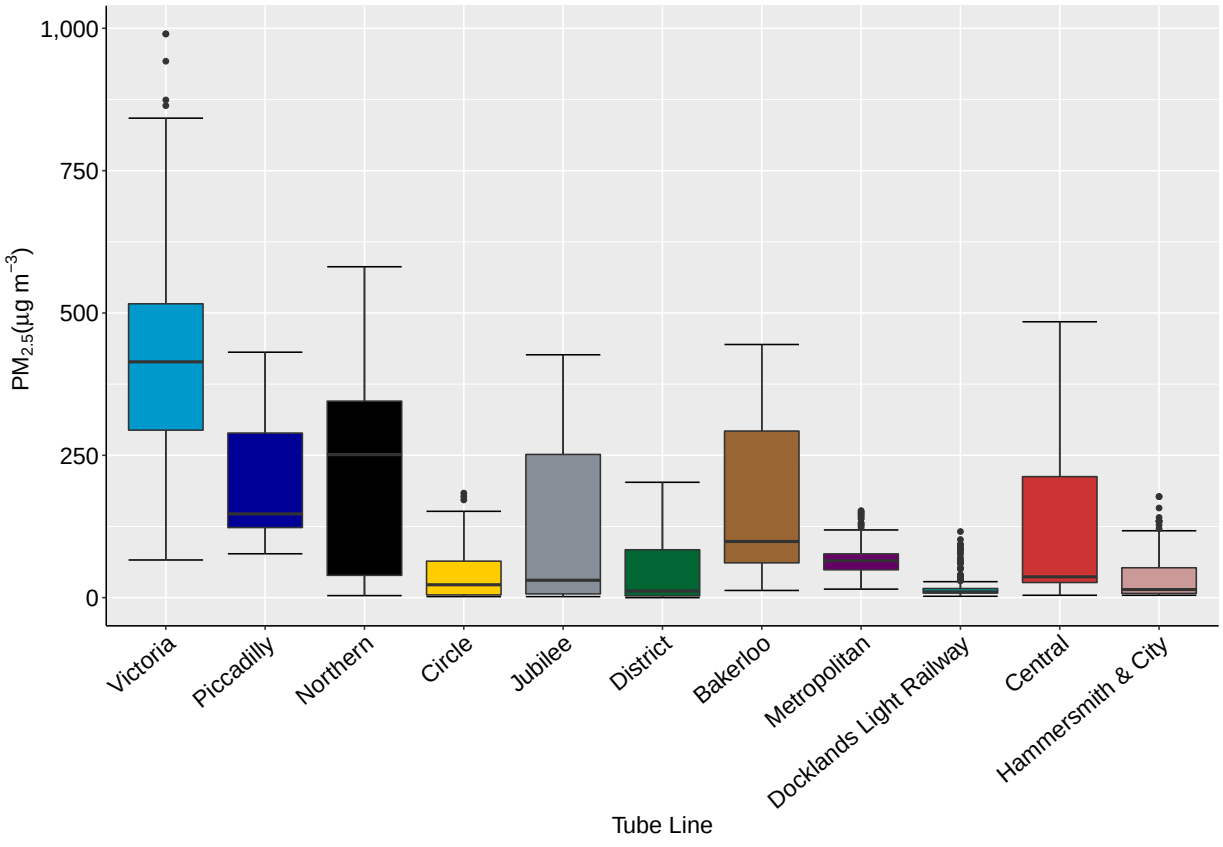


Figure 5.7: $PM_{2.5}$ μg m^{-3} on the tube, summarised by line. The lower and upper hinges correspond to the 25th and 75th percentiles, the horizontal line to the median, and the whiskers to 1.5 x the inter-quartile-range (approx. 95% percentile).

This visualisation gives a much clearer picture of the concentrations and variations found between lines, and although the boxplot has identified a number of concentrations on each

line as outliers, it is worth noting that this is likely not representative of bad data or device error, it is actually that there are a number of places on some of the lines with substantially higher concentrations that the median. This is most apparent with the Victoria Line where there are concentrations recorded of over 900 μg m^{-3} compared to the median of around 280 μg m^{-3}. Interestingly the Circle, District, Metropolitan, Docklands Light Railway, and Hammersmith & City lines all have noticeably lower medians than the other lines and a relatively small inter-quartile range. Given that the environment of the tube varies between fully exposed to the outside air (in the manner of a normal train), to semi-covered, to fully underground, it seems a reasonable premise that these different environments are effecting the rises and falls in concentrations as per the timelines in Figure 5.5. This was tested in Section 5.5.3.

5.5.3 Concentrations v. Depth

To compare concentrations by line, the mean station depth of all stations on a line was calculated and plotted against the mean $PM_{2.5}$ concentration for that line (Figure 5.8)

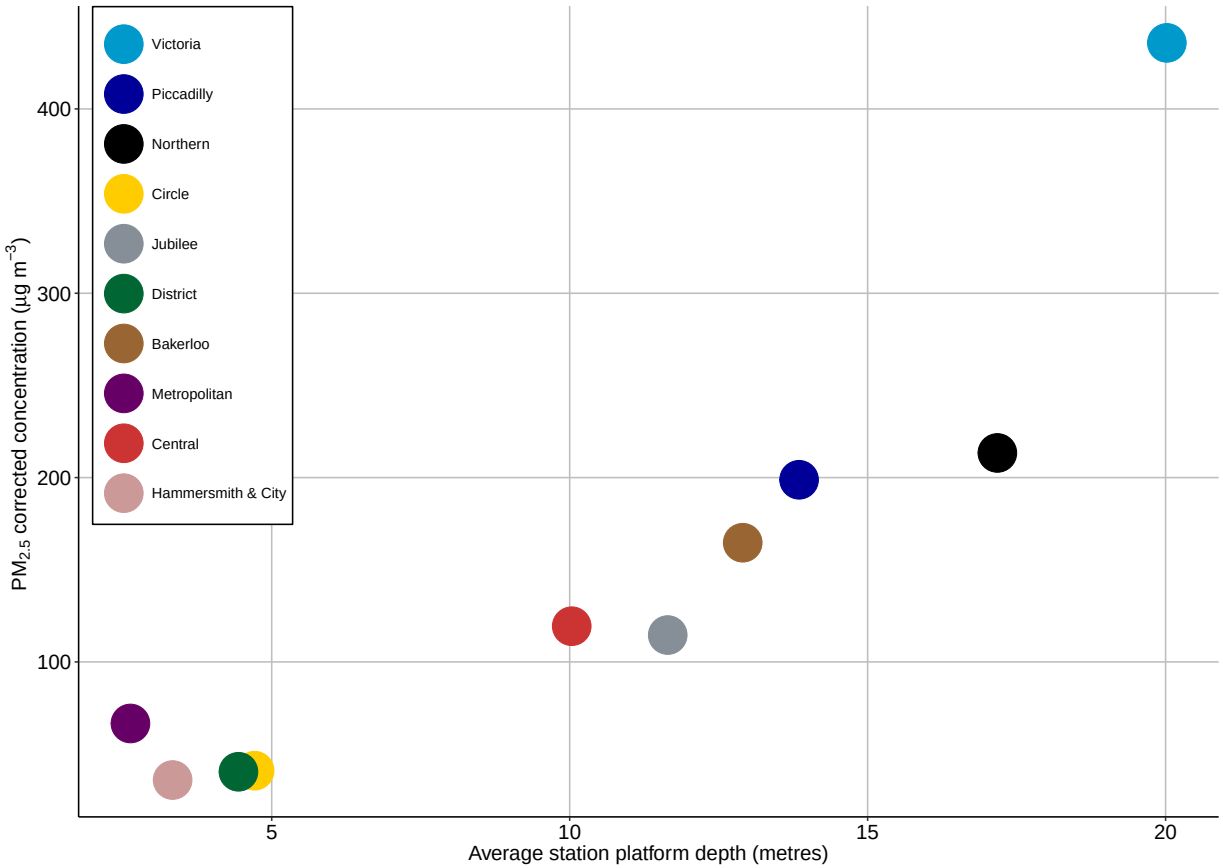


Figure 5.8: Mean concentrations v. mean station depths by tube line

Plotting the average line depths against average line concentrations appears to back-up that depth is an important determinant of levels of $PM_{2.5}$. The shallower lines have the lowest average concentrations, and the deepest lines the highest. To explore this further, figure 5.9 plots the concentrations for each station of each line in the same manner. Each point is the mean of the $PM_{2.5}$ concentrations recorded at that station.

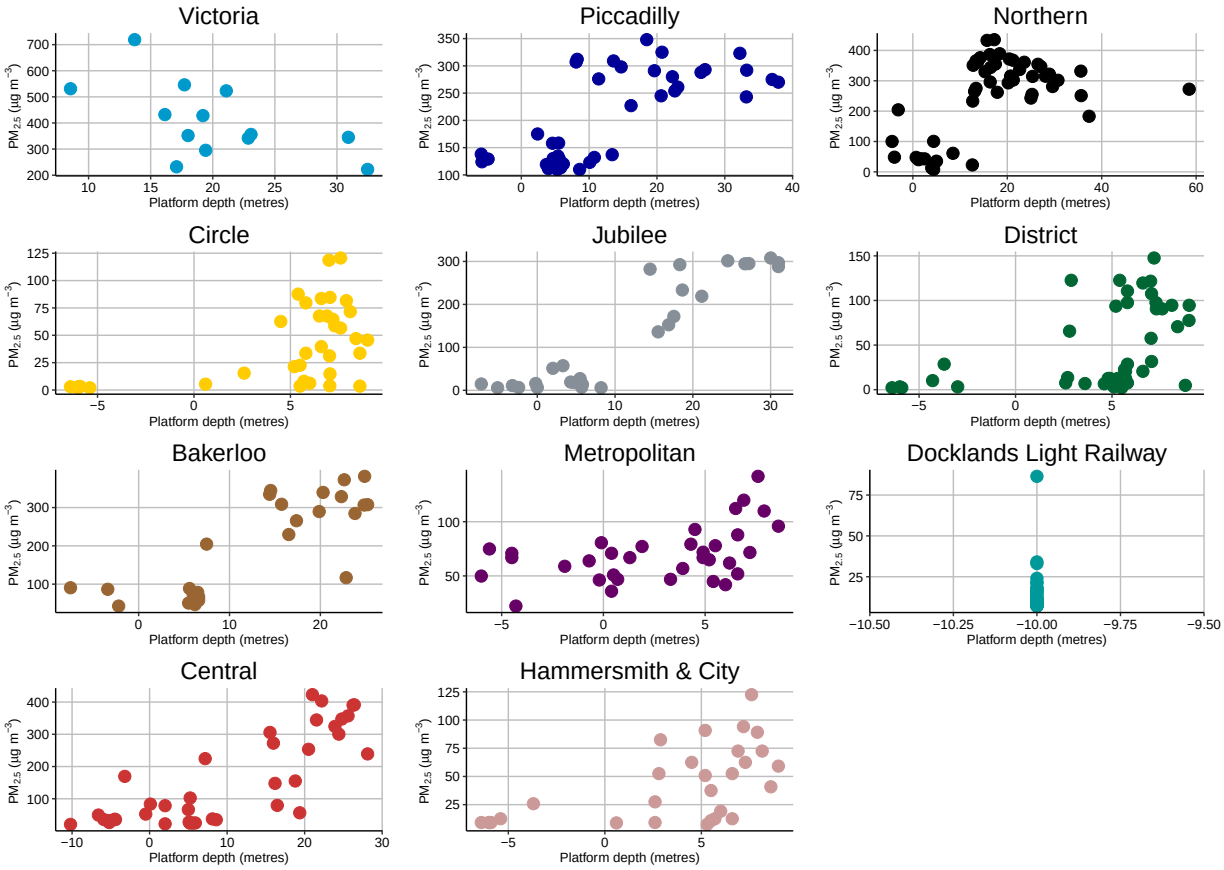


Figure 5.9: Concentrations recorded at stations v. station depth

The relationship between depth and concentrations that was apparent in Figure 5.8 is not so clear in Figure 5.9, i.e., once the depth and concentrations are plotted for each individual station as opposed to a line average. Although there still seems to be a relationship between depth and $PM_{2.5}$, it does not seem to be such a straight-forward relationship as increasing depth equals increasing concentrations. Taking the District line as an example (Figure 5.10), the stations mostly either have a depth of 5 metres above ground, or 5 metres below ground. The stations above ground (< 0 metres) tend to have low concentrations, but the stations below ground (> 0 metres) have both high and low concentrations (highlighted with black and blue boxes below in Figure 5.10).

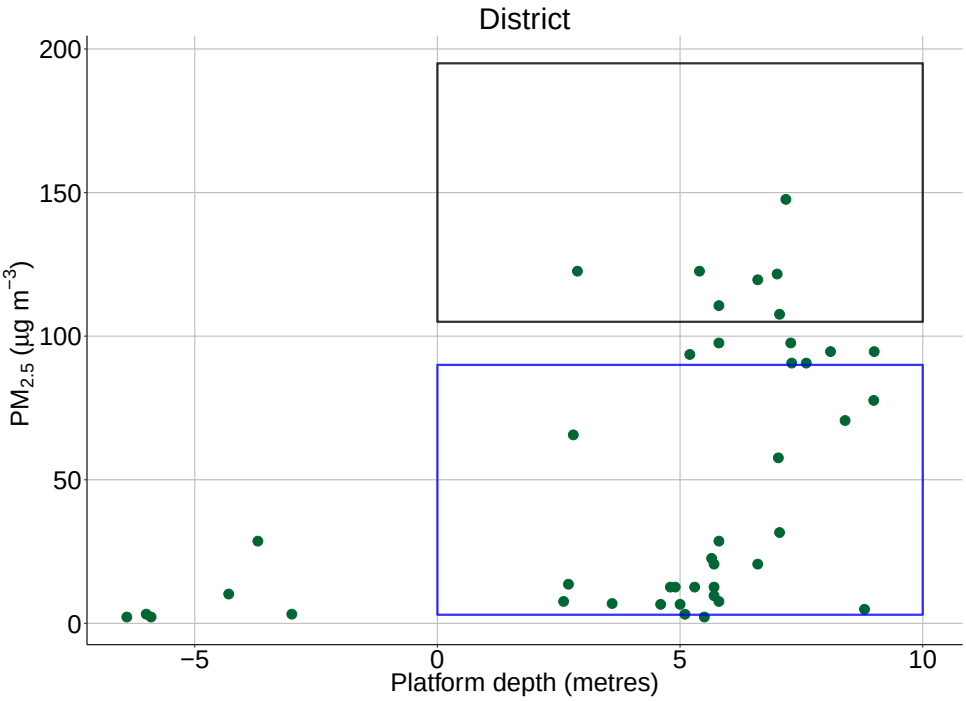


Figure 5.10: Concentrations recorded at District line stations v. station depth

Similarly taking the Central line (Figure 5.11), there is a general increase in $PM_{2.5}$ as station depth increases, except for Gants Hill and Wanstead. These two stations have a depth of 18-20 metres, but concentrations of only 100 μg m^{-3}, unlike other stations with similar depths where $PM_{2.5}$ is in the range 300-500 μg m^{-3} (marked with a black box).

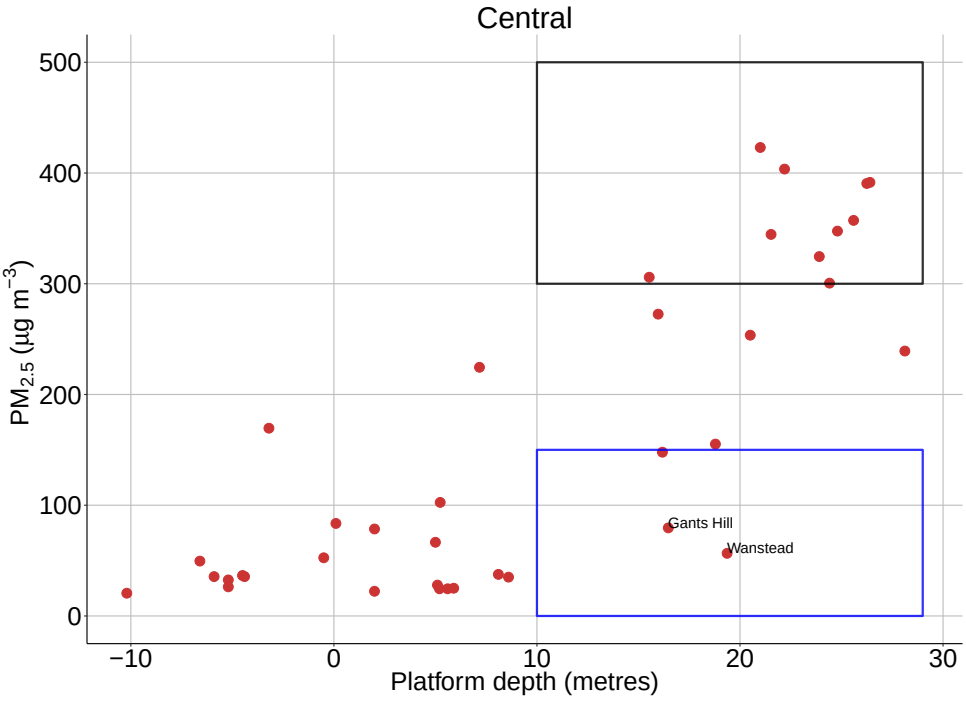


Figure 5.11: Concentrations recorded at Central line stations v. station depth

Manually inspecting the depths of Wanstead and Gants Hill, and the stations before and after them on the line, it becomes clear that these stations are in an area of the Central line network where stations are mostly shallow, and that these two are an exception. This could mean that shallow stations tend to be more well ventilated due to natural air circulation, and that this cleaner air is being moved down the tunnels to Gants Hill and Wanstead by the movement of the train, or indeed that the train cabin is being 'flushed' with cleaner air at those station platforms and that it does not build back up to higher concentrations by only going one stop. A plot of the geography of the line, along with concentrations and depth was made in Figure 5.12 to better understand this point (Code for creation in A.5).

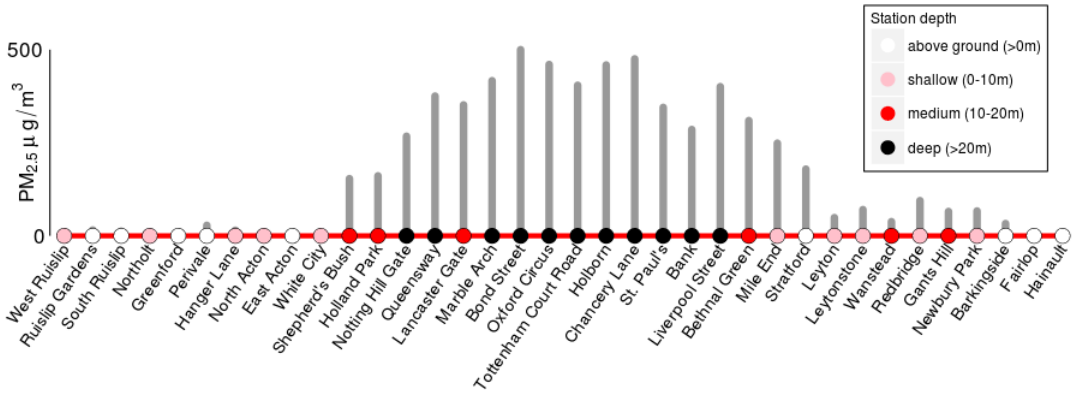


Figure 5.12: Central line locations, depths and $PM_{2.5}$

From the plot we can see that Wanstead and Gants Hill (towards the right of the graph) are both medium depth stations (shown as red circles), and as such might be expected to have pollutant concentrations similar to other medium depth stations such as Shepherd's Busy, Holland Park, Bethnal Green and Lancaster Gate. But the $PM_{2.5}$ levels are actually more like that of above ground and shallow stations, suggesting that the depth of a station is not directly related to $PM_{2.5}$ concentrations, and that distance from outdoors or shallower stations is important too.

Taking a further example of stations and concentrations that do not seem to fit the pattern of deep equals high, and shallow equals low, Golders Green has a depth of less than 0, i.e. above ground, but concentrations of over 200 μg m^{-3}. From manual inspection of the data, this station is located on the North-West spur of the Northern end of the Northern line. The mean value for the station of 200 μg m^{-3} is calculated as an average of four four recorded values; 356 μg m^{-3}, 14 μg m^{-3}, 374 μg m^{-3} and 74 μg m^{-3}. Looking at this data in more detail, the higher two recorded concentrations relate to the train arriving at Golders Green having just emerged from the tunnel and deep station of Hampstead,

and the lower concentrations are when the train has arrived from East Finchley (an above ground station further North on the track). The difference in these values is quite extreme and suggests that for stations that are outside or shallow, but close to deeper stations, the direction of travel influences the $PM_{2.5}$ concentrations for that location. As the tube train arrives at Hampstead from within the deep tunnels further South, the carriage must still be full of air that has accumulated over the journey, which will get partly flushed out by the doors opening at Golders Green, but not before the Sidepak device (with one minute resolution) has recorded high concentrations 'at' Golders Green, which then drop by the next station at East Finchley. Conversely, when the tube arrives at Golders Green from East Finchley, the air inside the carriage is much cleaner as it has not been deep underground previously. The concentrations inside the carriage only start to rise to levels of 200 - 400 μg m^{-3} once tube has gone into the tunnel at Hampstead and started to be exposed to the higher concentrations found in those deeper tunnels.

To further explore this pattern, the data for Oxford Circus on the Victoria line was examined. There were four measurements taken while on the train at that station, which were 140 μg m^{-3}, 322 μg m^{-3}, 152 μg m^{-3} and 362 μg m^{-3}. The lower two measurements (140 and 152) coincide with the train heading Northbound, and Southbound respectively. Direction of travel seeming to have have very little effect on the concentrations at this station. Though as this whole are of Victoria Line track is deep and quite a way from any outdoor stations, this seems to follow.

In summary, from this small sample, it seems that $PM_{2.5}$ concentrations in the carriage at stations which are underground, and are not close to a platform or section of line that is outdoor, are not effecting by the direction of travel. Conversely, those that are near outdoor sections of line and platforms, are. The effect of this variation is partly minimised due to our study design, i.e. we measured each station arriving and departing from different directions.

5.5.4 Spatial distribution of tube air quality

Figure 5.13 shows a screenshot of an online map of tube stations, with the colour of the station name used to indicate the mean levels of $PM_{2.5}$ recorded (in conjunction with the scale on the right). The map is available at: `http://londonair.org.uk/modeling/tube-pm25/map.html`.

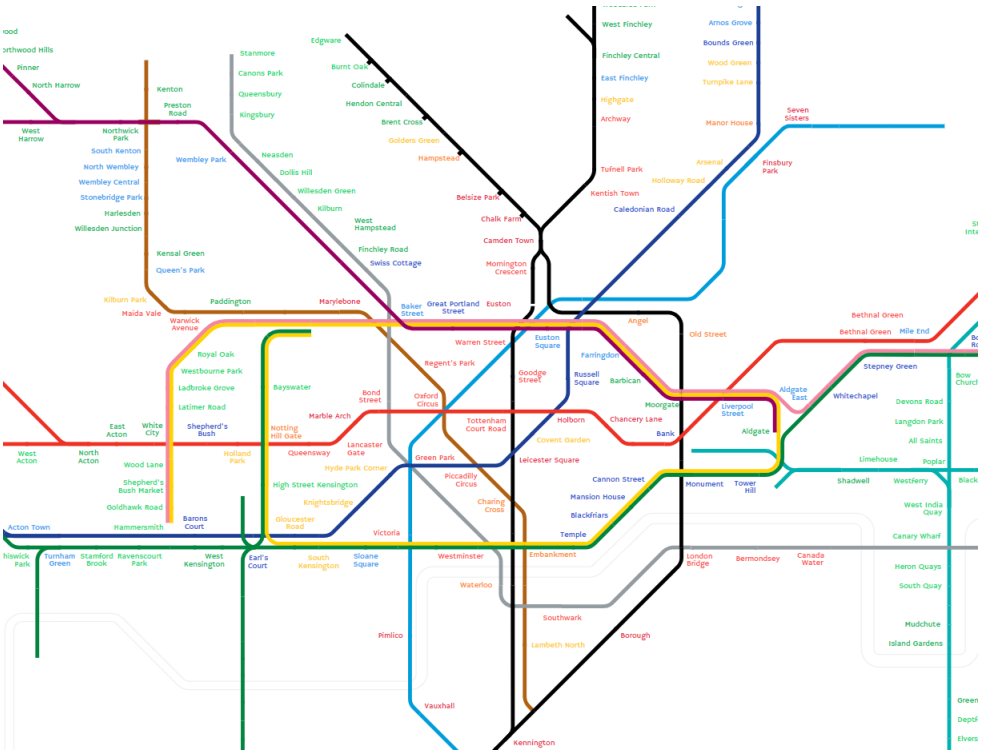


Figure 5.13: Locations of tube stations and $PM_{2.5}$ levels

When using the map and interacting with the data, as noted previously, there looks to be a relationship between depth and $PM_{2.5}$ concentrations i.e. stations in central London have higher concentrations, due to the lines being deeper and more frequently underground than in outer London. However as noted in subsection 5.5.3 taking means at stations can hide variation, so whilst this map is useful for giving a general impression of the spatial variation of concentrations, a more sophisticated approach might be more useful for modelling exposure.

5.5.5 PM build-up and dissipation

In Figure 5.5, where timelines of $PM_{2.5}$ are shown for each tube line, it is possible to see how on certain lines and at certain places concentrations fall to levels similar to background concentrations. To investigate how quickly the air in the carriage falls to these levels, from the elevated levels, the Jubilee line was taken as an example (due to it's clear differences between high and low concentrations). Figure 5.5 has been re-created and annotated as Figure 5.14 below; a red line has been added to show the background $PM_{2.5}$ concentration (taken from the 'Kensington and Chelsea - North Ken' background monitoring site), and black boxes and station names have been added to clarify the areas of interest.

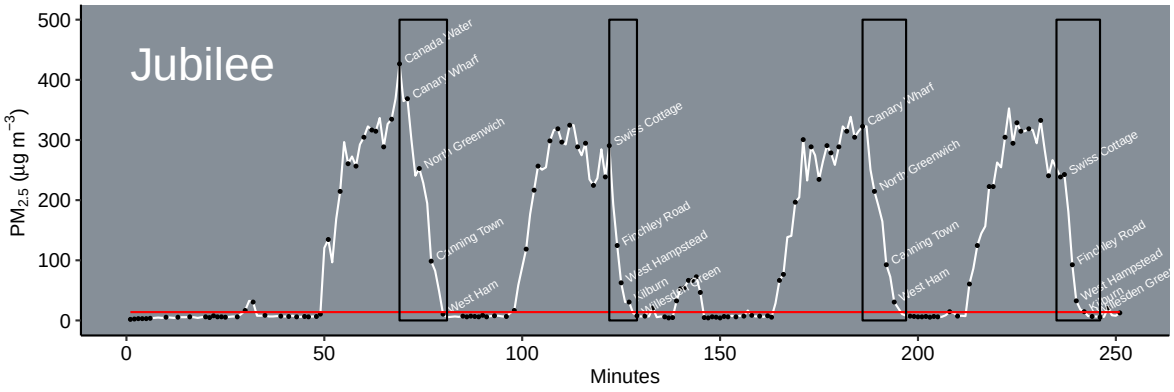


Figure 5.14: Locations of tube stations and $PM_{2.5}$ levels

Taking the first and third highlighted sections above, concentrations build-up inside the cabin while the train is underground, but then start to fall after Canada Water, reaching background levels by the time the train is at West Ham. The depth of these stations are, in order, Canada Water (18m), Canary Wharf (18m), North Greenwich (15m), Canning Town (2), and West Ham (0m). These stations do not have the step-by-step change in depth in the way that the concentrations do. If anything, the first three stations in this subset might be grouped as deep, and then the final two as surface. However the concentrations do not immediately change from high to background, they gradually decline. This suggests that in addition to depth being an indicator of concentrations inside the tube trains, distance from cleaner outside air, and it's exchange with air inside the cabin when the doors open, also effects concentrations. To elaborate, the concentrations between Canada Water and North Greenwich fall by about 40%, despite there only being a small change in depth. This suggests that the change in concentrations is due to the doors opening at North Greenwich, and an air exchange happening with the air on the platform, which is cleaner 'platform air' than is the case at Canada Water, due to North Greenwich being closer to a surface station (Canning Town). Now taking the second and fourth sections highlighted in Figure 5.14 the stations under consideration are Swiss Cottage (17m), Finchley Road (3m), West Hampstead (5m), Kilburn (7m) and Willesden Green (5m). Here we see a similar pattern, in that the concentrations do not immediately drop to background levels in the way that might be expected if depth was the sole determinant. It takes a couple of stations (and subsequent doors opening and air exchange) for the air in the train cabin to be sufficiently replaced with cleaner outside air.

5.5.6 Train stock

Mean concentrations per line were now calculated (in the same manner as section 5.5.2), but rather than grouping by line, the train stock data was linked, and boxplots created with that as the categorical variable (Figure 5.15)

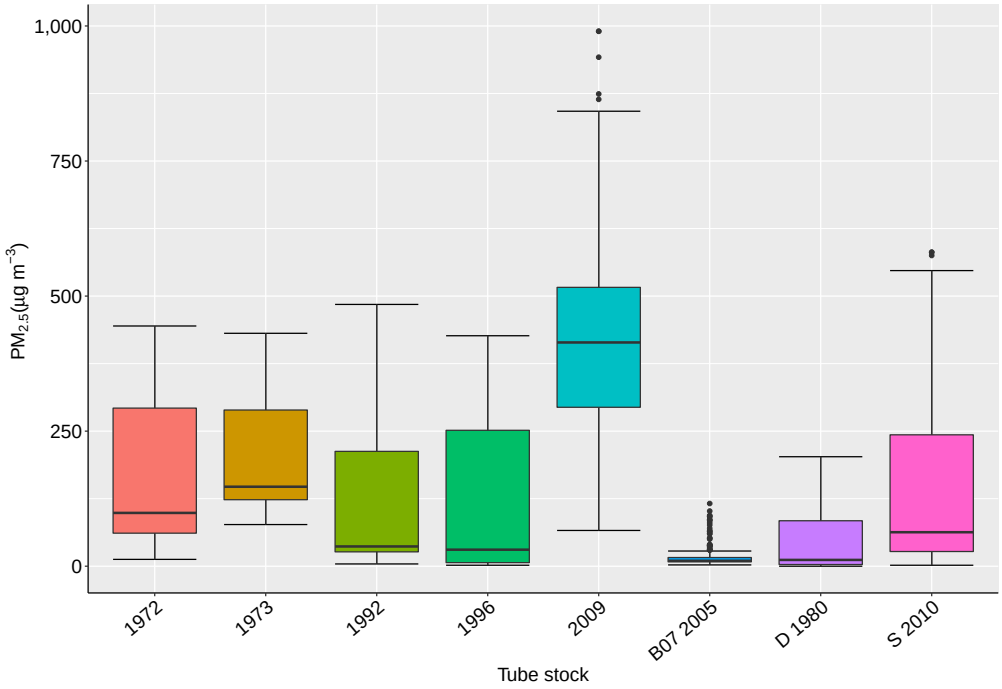


Figure 5.15: $PM_{2.5}$ μg m^{-3} on the tube, summarised by stock type

There does not seem to be any clear pattern to $PM_{2.5}$ concentrations when examining the data using train stock as a variable. Excluding the B07 2005 stock (as they are soley used on the DLR), the oldest trains (D 1980 stock) have the lowest concentrations, and the 2009 stock the highest. But they are only used on the District line and the Victoria lines respectively, and as we have seen there are large variations within the length of those lines which suggest that there are other factors (namely depth and distance from exposed station platforms) which are influential and not linked to train stock.

5.5.7 Revising LHEM exposure estimates

As discussed at the beginning of this chapter, whilst travelling on the London Underground network the LTDS-X subjects were assigned $PM_{2.5}$ exposure concentrations of 95 μg m^{-3} per minute. We can now see that this is a simplistic representation of the exposure found within the network. A final aim for this research area is to create a detailed spatial model layer

which can be used for exposure assessments of the population of London while on the tube, however as an intermediary step, two LTDS subjects who used the London Underground during their day were chosen at random, and their exposure recalculated, but with mean line concentrations taken from this new data, to give an example of the possible effects using it will have. Their LHEM exposure (baseline in this instance) before this new method is shown in Figure 5.16, with the exposure of 95 μg m^{-3} while the subjects are on the tube, shown in blue.

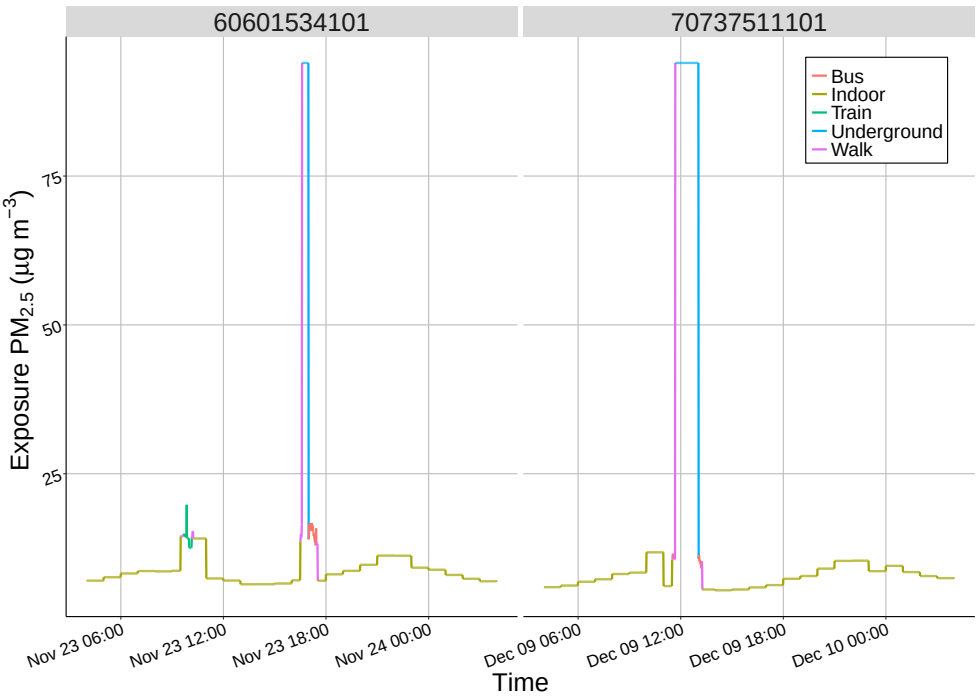


Figure 5.16: PM$_{2.5}$ μg m^{-3} exposure for LHEM subjects 60601534101 and 70737511101

The mean daily exposure for these two subjects was 10.06 μg m^{-3} and 12.69 μg m^{-3} respectively. Subject 60601534101 begun their journey at Finsbury Park, and ended it at Arnos Grove, taking the Piccadilly line. The mean PM$_{2.5}$ recorded on the Piccadilly line was 63 μg m^{-3}, and this is therefore substituted instead of the 95 μg m^{-3} used. For subject 70737511101, they began their journey at Rayners Lane, and ended it Canning Town, taking the Metropolitan line for approximately the first third of their journey, and the Jubilee line for the remaining two thirds. So the first third of their journey the mean Metropolitan line concentration of 67 μg m^{-3} is used, and for the second two thirds of their journey the mean Jubilee line concentration of 138 μg m^{-3} is used. The new timelines are shown below in Figure 5.17.

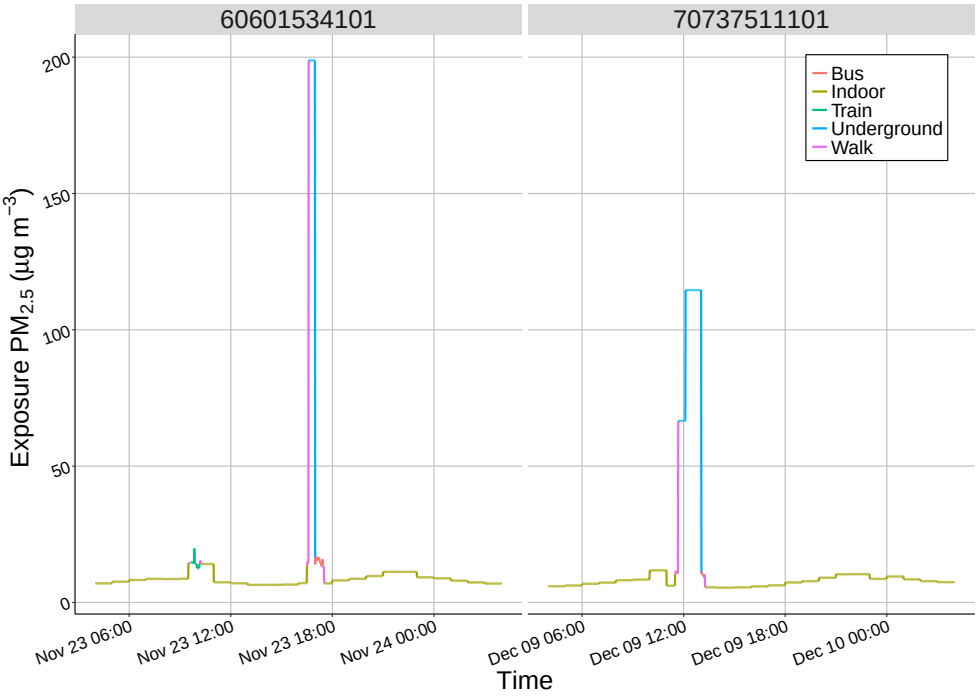


Figure 5.17: Revised $PM_{2.5}$ μg m^{-3} exposure for LHEM subjects 60601534101 and 70737511101

The new mean daily exposure for these two subjects, using the new London Underground data, was 11.73 μg m^{-3} and 12.99 μg m^{-3}, an increase of 17% and 2.4% respectively. The former being higher, as although the journey was short the concentrations were much higher for that period than previously modelled. Alongside this increase in mean exposure, both subjects for a short period are also exposed to higher peaks in concentrations than in the old method.

5.6 Discussion

This dataset is, to the best of my knowledge, the largest, most systematic, and detailed collection of $PM_{2.5}$ concentrations on the London Underground. There are few studies which have studied $PM_{2.5}$ on the tube, however in a review of exposure on metro systems Nieuwenhuijsen et al. (2007) collated various studies, and specifically within London found ranges of between 130–200 μg m^{-3}, 157–247 μg m^{-3}, and 12–264 μg m^{-3}, mostly from the Adams et al. (2001b) studies which completed around sixty journeys on the tube in 2001. The data collected for this chapter is superior as the Adams work only collected data for short sections of repeat journeys on the Piccadilly, Bakerloo, Northern and District lines for comparison with other transport modes, and no spatial analysis was undertaken (or

indeed possible as the data was collected on filters and therefore lacked spatial and temporal granularity).

In addition to peer-reviewed academic work, the main other sources of $PM_{2.5}$ measurements on the tube come from occupational health work commissioned by TfL, the most well know being the "Assessment of health effects of long-term occupational exposure to tunnel dust in the London Underground" led by Hurley and published in 2003 (Hurley et al. (2003)). This was commissioned with the purpose of looking at occupational exposure to drivers and station staff, with only a small section concerning passengers, and therefore most of the results focus on the personal exposure of drivers and staff. There is no consideration of the variability by line, or within line, or an attempt at understanding the viability. In summary, the work completed and data collected is appropriate for for the purpose it was commissioned for, but not there are no spatial or temporal attributes linked to the data to enable further investigation or to use the results in other ways. Nonetheless the $PM_{2.5}$ levels were found to be in the range 270–480 μg m^{-3} on station platforms, and passengers average expose was taken as 200 μg m^{-3}, although a number of broad assumptions were made to arrive at this figure. As with Nieuwenhuijsen et al. (2007), these concentrations are not dissimilar to those we measured.

By collecting $PM_{2.5}$ concentrations and linking them to time-resolved location data across the whole of the network, and further calibrating the $PM_{2.5}$ measurements using newly calculated scaling factors, we have been able to offer the most complete understanding of the variation and levels of pollutant in this environment that is used by millions of people everyday. There are however some issues and considerations with the methods and data collection that need discussion, and which might effect our findings.

Firstly, all of our measurements are taken from within the carriage of the tube. So whilst there is often discussion of stations within this chapter, this is actually the tube train (normally with doors open) pausing at the platform of a station for a minute or two before moving away again, and the reader must be careful not to misinterpret the findings as such.

Another area that might warrant further refinement is around the effect that passenger numbers have on concentrations. The theory of the 'personal cloud', that being that a persons movements and activity can stir-up particles into the air that otherwise may have settled on surfaces. With so many people moving inside trains and on the platforms of the stations, this might contribute to increased concentrations independently of other factors i.e. the movement of the trains. This said, studies (Ferro et al. (2004)) have found that this personal cloud effect to only increase concentrations by a few μg m^{-3}, which when set alongside the 100s of μg m^{-3} being recorded are quite insignificant, and thus it seems that

the effect of passenger numbers can be largely ignored in efforts to understand $PM_{2.5}$ levels in this environment.

Similarly, Adams et al. (2001b) discusses that wind direction and outdoors concentrations may be linked to increased concentrations in the tube system, with particles being blown into the tunnels and recirculating. But with background concentrations of $PM_{2.5}$ in London normally around the 10–15 μg m^{-3} mark, it seems unlikely that pollutants from outside the system are contributing in a meaningful way to the high concentrations found (when the tube is in the underground environment anyway).

Regarding the findings related to depth, these must also be considered alongside the fact that the depth data obtained and used in this research related to station platform depths, and was not detailed enough to enable understanding of depth between stations. This would have been useful particular for Section 5.5.5 (PM build-up and dissipation) where the build-up and dissipation of $PM_{2.5}$ between stations was considered. Alongside monitoring equipment with a higher time resolution, say 3-4 seconds rather than 1 minute, it would be easier to understand how the concentrations vary between stations and along stretches of track.

A further complication to the findings we have so far, is that of ventilation settings and train stock. Some of the lines only have one stock-type running on them, but some lines have a variety. Also within these varieties different ventilation settings are available, and vary as to whether they are in use/functioning or not. It would be useful to repeat a sample of this data collected with a specific focus on ventilation to understand this variable more.

The data collected in this study is an important step forward in better estimating the exposure of people who live and commute in London (as was demonstrated with two random LHEM subjects). With further work, a dataset can be created which will allow vastly improved estimation of the exposure of millions of people travelling on the tube, which in most exposure studies involving London is not currently considered at all. Epidemiological studies are still geocoding peoples address or perhaps postcodes, and then taking annual concentration maps, normally with small ranges between the maximum and minimums, when there are millions of people every day who are spending prolonged periods of time in environments that have $PM_{2.5}$ concentrations of over 500 μg m^{-3}. It might be that the there are strong links between people who use the London Underground for more than an hour a day, and those who develop chronic obstructive pulmonary disorder (COPD) issues. At the moment there has been no way to investigate this on a population level, but this research makes developing a research question about this subject now possible.

Future work in this area should focus on creation of a geographically defined data layer or database which can be used alongside passenger modelling (such as done by a Dr Reades,

a colleague in the Geography department of KCL (`https://www.youtube.com/watch?v=F6qsh1KBW-E`).

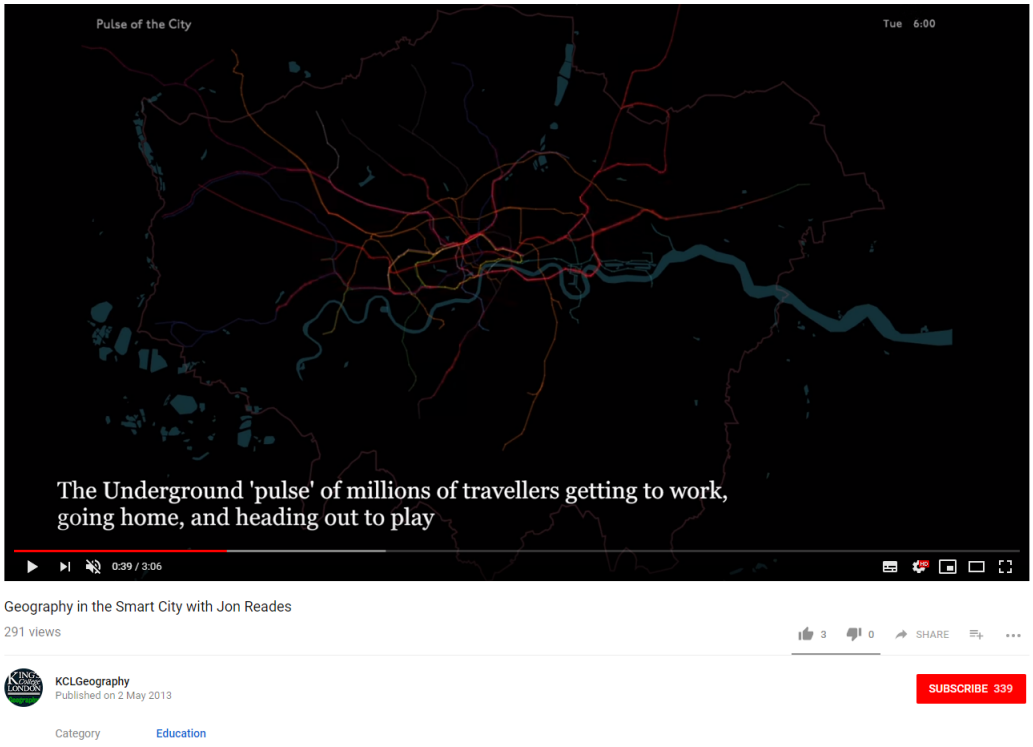


Figure 5.18: Dr Reades explaining the 'pulse of London' through geographical data, explaining how travel data can be used to investigate wider sociological questions

This dataset should be sufficiently geographically defined, to have different exposure estimates related to each section of each tube line, rather than the intermediate step used in Section 5.5.7 whereby line means were calculated. Preferably it should also take into account the direction of travel, as this was seen to alter $PM_{2.5}$ exposure dramatically. One option would be to create a Tube network within a pgRouting database, and assign the concentrations as 'cost parameters' to each stretch of the network. Once built this would allow input of an origin and destination, and then the algorithms would calculate a route between the two stations that minimised exposure, with an output of exposure along the route. Although the LHEM currently makes use of the TfL routing API for this step of the model, and therefore a layer that can be easily incorporated to this step would be a more appropriate approach. Each minute that an individual is on the Tube would be a location and timestamp, and also a field containing the name of the Tube line the journey is on, so these coordinates could be 'snapped' to the nearest Tube segment, for the appropriate line of travel, and then the relevant concentrations extracted from the line data and assigned to that minute of travel.

5.7 Conclusions

$PM_{2.5}$ data collected on the London Underground varies between 0 μg m^{-3} and 990 μg m^{-3}, with a mean of 129 μg m^{-3} and a median of 63 μg m^{-3}.

There is a large variation when comparing lines against each other, and between sections of the same line. For example the Victoria line has concentrations of 990 μg m^{-3} in some places, but a mean of 436 μg m^{-3} and a minimum of 66 μg m^{-3}.

When ranked in decreasing order of mean $PM_{2.5}$ concentrations, the results are; Victoria Line (436 μg m^{-3}), Northern (219 μg m^{-3}), Piccadilly (199 μg m^{-3}), Bakerloo (164 μg m^{-3}), Central (119 μg m^{-3}), Jubilee (115 μg m^{-3}), Metropolitan (67 μg m^{-3}), Circle (41 μg m^{-3}), District (40 μg m^{-3}), Hammersmith & City (36 μg m^{-3}), Docklands Light Railway (18 μg m^{-3}) . Stations and sections of lines also have large variation. The Southern section of the Victoria line had some of the highest concentrations, particularly on the stretch of track between Brixton, Stockwell and Victoria with concentrations of 400-500 μg m^{-3}, compared to many stretches of track on more exposed lines such as the Circle, Hammersmith & City and the Metropolitan lines where concentrations are often less than 10 μg m^{-3}

Increasing depth was considered to be an indicator of increasing $PM_{2.5}$ concentrations, illustrated by the linear relationships seen in Figure 5.8 which looked at average concentrations by line and plotted them against average depth. However when this plot was dis-aggregated and considered by individual line, station, and direction of travel, this relationship became more complex. There were high concentrations recorded while on the tube at stations that are not particularly deep, and conversely low concentrations recorded while on the tube at stations that are very deep. The explanation for this variation is due to the environments immediately previously experienced by the train, and distance from those environments. Trains that have just exited areas of high concentrations bring polluted air with them which takes time to dissipate, and conversely tube trains that have been in cleaner air take a little while for concentrations in the carriages to increase. This increase and decrease seems likely to mainly happen at platforms when doors are opened and closed, but also to a lesser degree during movement of the train (both of which need further investigation).

The spatial representation of the stations in London, ignorant of direction of travel and depth data, showed that higher concentrations tend to be found in central London areas.

Finally, by calculating simplistic line averages in lieu of a more complex dataset that will be created in the future, two randomly chosen LHEM daily exposure estimates were recalculated. Both subjects daily exposure increased, one by 17% and one by 2%. Repeating

this new method across the entire tube-using population of London is likely to lead to a general increase in exposure from the LHEM, but with some people experiencing lower overall exposure (due to only travelling on section of the tube that are cleaner than previously presumed).

6. Evaluating dynamic exposure models

6.1 Aim

Develop an understanding of methods to evaluate predictions of exposure from hybrid-type models

6.2 Objectives

- Develop a data collection plan based on simulated and measured datasets
- Undertake mobile monitoring to collect data representative journey(s)
- Model exposure of the same journey(s)
- Analysis: compare the monitored and modelled exposures

6.3 Background

The focus of this PhD research so far has been on developing a dynamic exposure model to better understand exposure to urban air pollution in the population of London. Having reconstructed the time-activity of the population, their exposure to $PM_{2.5}$ and NO_2 was modelled, and then refined, with further investigation of the London Underground micro-environment. This next chapter will consider how to evaluate the NO_2 results that are calculated using an exposure model of this style.

The features that a hybrid exposure model should have were defined in Section 2.4.4 as *"It should have highly temporal and spatially resolved air quality inputs which consider both indoor and outdoor sources (including regional and local source for the latter), it should be able to model infiltration rates for different modes of transport and building types, it should*

reflect the multiple micro-environments that people spend their time in (and take account of the temporal resolution of these) and finally it should (for linkage through to epidemiological end-points) be able to consider different breathing rates to quantify exposure and dose for multiple pollutants". Section 2.4.4 examined models that were within this wider field (of varying levels of complexity), but it was noticeable that there was little evaluation of the exposure predictions that they made. There are to my knowledge no established protocols for evaluating exposure predictions from a hybrid/dynamic exposure model, as this type of method and field of research is relatively new. In addition, as the field grows the exact approaches are being refined and vary between studies, meaning one evaluation method would unlikely be fit for the next study. Possible sources of error in this type of model are classified in Table 6.1 (below), with a brief description, and whether they are unique to exposure models of this kind.

Table 6.1: Errors and uncertainty in exposure models

Type of error	**Description**	**Hybrid specific**
Air quality annual average monitoring site predictions	Evaluation exercises of air quality models using high quality monitoring site data demonstrate that they (in our case CMAQ-UK) can make predictions of annual averages of most pollutants with reasonable accuracy.	No, CMAQ-UK and other similar models are commonly used in static exposure studies.
Predicting annual average concentrations at sites without monitoring data	How air quality models predict concentrations in locations that are not readily available for evaluation via monitoring sites is not well understood.	No, CMAQ-UK and other similar models are commonly used in static exposure studies.
Air quality temporal resolution	Annual averages are often used for exposure studies, which are relatively easy to quantify against monitoring sites (see above). However the complexity of the hybrid model we are now considering uses hourly diurnal profiles, which vary in their accuracy of prediction over time, and therefore the accuracy of their input to exposure varies by time of the day, and day of the week.	Partly. Not many static studies of large groups of people consider exposure at high temporal resolution (because they normally use annual averages), and even less quantify and include the errors that this produces.
Micro-environmental modelling	Mass-balance models and I/O ratios to estimate the concentrations of pollutants within microenvironments, in relation to outdoor concentrations, are inherently subject to variation due to the inexact inputs. Literature reviews to establish best-guess I/O ratios are common but their transfer-ability to other counties/cities is often unknown. Monte-Carlo simulation of the inputs to create a range of predictions can be undertaken to understand their impact.	Yes. Static exposure studies normally use outdoor concentrations for exposure assessment. Exposure assessments of the health effects of environments i.e. indoor, may use micro-environmental modelling and I/O ratios, but not in conjunction with peoples movements, multiple environments, and high spatial-temporal air quality.
Temporal representative errors of exposure	Exposure predictions for a person over a time period can be made, but the degree to which these predictions are representative of that persons exposure for that period of time are unknown. In the LHEM the exposures were calculated based on the persons previous days movements, and the respondents were asked whether this was representative of their typical day; but quantifying the difference between the day of the data collected and their typical day in terms of exposure is not explored.	Yes. Hybrid exposure studies that consider the exposure of individuals through space-time, especially those that seek to frame exposure results in terms of longer-term health effects, need to develop methods to estimate the variability of representativeness error.
Representative errors of groups and populations	Extrapolating exposure predictions from a small group of people (e.g. a classroom of 30 ten year olds in South London), to larger groups of people (e.g. all school children in London) can be controlled in part by statistical sampling techniques and appropriate power calculations. But this can only be done with prior data/knowledge of the population, which while fairly simple to do for basic demographics using Census data and similar, is much more difficult to do for exposure prediction models. To ensure representativeness exposure sample calculations need to ensure that important drivers of exposure i.e. tube usage, are included alongside more standard variables.	No, other exposure methods have this issue, but it is often not addressed fully. Studies with larger numbers of participants would be expected to have a better range of exposures and be more representative.

The next sections of the background to this chapter are discussed around the types of uncertainty from Table 6.1 above.

6.3.1 Air quality annual average monitoring site predictions

Air quality models used in hybrid exposure studies are often evaluated against annual averages from monitoring sites in the cities of the study. For example the study by de Nazelle et al. (2013) in Barcelona used a dispersion model (Lao et al. (2011)) for the NO_2 air quality input to their exposure model, and evaluated it against a number of monitoring sites, demonstrating good performance/agreement (Table 6.1) with errors in the range of -12% to +10%.

Barcelona monitoring sites	Type of location	Actual NO_2 ($\mu g/m^3$)	Modelled NO_2 ($\mu g/m^3$)	Model / Actual (%)
Ciutadella	Urban background	42.3	46.2	109%
Vall d'Hebrón	Urban background	36.5	37.7	103%
Eixample	High traffic site	65.4	63.2	97%
Gràcia	High traffic site	62.6	57.9	93%
Poblenou	Moderate traffic site	47.4	41.8	88%
Sants	Moderate traffic site	45.3	50.0	110%
Average value	---	49.9	49.5	99%

Figure 6.1: Performance statistics of an air quality model used in de Nazelle et al. (2013)

By understanding the scale of these errors in air quality models, and by presuming that they are uniform in time and space across the study area (discussed more below), they can be incorporated into a static exposure study of outdoor air at household addresses or similar.

6.3.2 Air quality annual average non-monitoring site predictions

The predictions of air quality models, away from monitoring sites, is less well understood. Taking a NO_2 CMAQ-UK model of London (Figure 6.2 below) we can see the location of monitoring sites i.e. where the predictions have been evaluated and understood. Clearly there are many areas which are not evaluated; there are 16 million cells in the raster, and only 242 sites (many of which are either not operational, do not have a high enough data capture rate for comparison, or do not measure the pollutants we are interested in, meaning the actual number is far less). This means that only 0.0015% of the locations shown in Figure 6.2 have been evaluated and the possible error (between model prediction and measurement) understood.

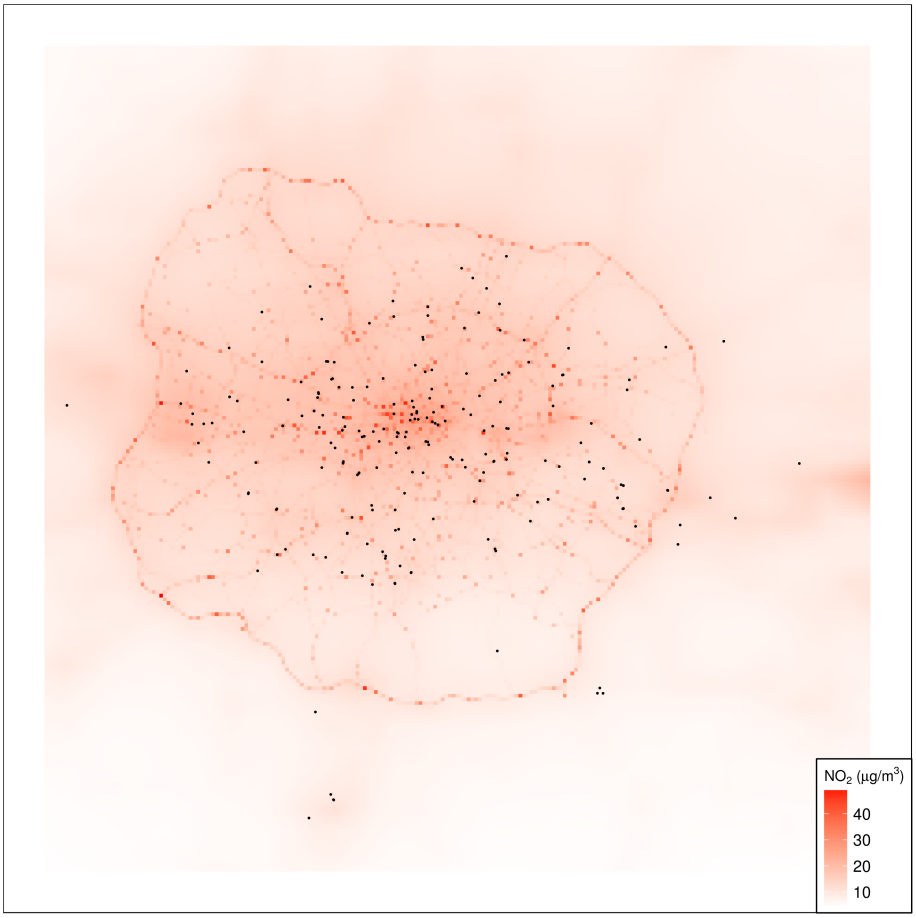


Figure 6.2: CMAQ-UK for London + Air Quality monitoring sites

For these non-monitoring site locations, we presume that the model functions in the same manner, with the same error levels. In reality it seems likely that there are varying degrees of accuracy. This may be less of a concern within an exposure model looking at health effects between areas of high/low concentrations, providing it represents the spatial pattern adequately i.e. predicting lower concentrations and higher concentrations in the right places, even if the concentrations themselves are not exactly right. As part of this chapter the aim is to better understand how the model predicts away from monitoring sites. As it is impractical to place a monitoring site at every location, a method is developed whereby mobile monitoring devices are used in-lieu of monitoring sites. As background to this, a review of the small number of studies which have already attempted this was undertaken. Buteau et al. (2017) compared a number of methods for predicting concentrations of O_3 and NO_2 in Montreal, Canada at the level three-character postal code area (6km^2 square). Alongside traditional spatial interpolation using inverse-distance weighting, and nearest monitoring site methods, they also based a land-use regression model upon a dense monitoring survey. However, these were semi-permanent monitors at fixed sites for 9 months of a year, and so whilst

their results were encouraging the method is not really transferable to portable monitors which are used here. Shi et al. (2016) also created a land-use regression model, with mobile measurements as an input, which achieved good results, but their study was based on repeated measurements using vehicles sampling along a set route for 14 consecutive days, and the choice of 14 rather than say 20 seemed to be a function of their available sampling time rather than calculated using statistical analysis.

6.3.3 Air quality temporal predictions

Air quality models that use a higher temporal resolution have more sources of error to quantify and include in an exposure model than one which uses annual averages. As an annual average value, a difference of $+/-$ 5 μg m^{-3} between the model and the measurements can be factored into the exposure prediction relatively easily. But if the model is predicting a concentration every hour then this error will vary for each hour i.e. 8760 hours resulting in 8760 model/measurement differences.

6.3.4 Micro-environmental modelling

Within dynamic exposure models, exposure in microenvironments is often calculated in relation to outdoor concentrations. Therefore, in addition to the uncertainty of the air quality predictions in a place and time, further uncertainty is added when the outdoor concentration is converted to a microenvironment concentration (whether this is done using an I/O ratio for indoor concentrations, or a mass-balance model as used in Section 4.4.1.3). Exposures in the LHEM, for all environments except the London Underground, used the CMAQ-UK air quality model as an input. Dhondt et al. (2012) described the possible sources of error within the in-vehicle section of their dynamic model, and noted that the air quality inputs may not be of a high enough spatial resolution to capture concentration gradients near roads adequately. But they did not seek to evaluate the individual-level daily/weekly/hourly concentrations, and indeed due to the type of the study (an aggregated population) they would have found it difficult to do so (due to the large numbers of individuals). Similarly and whilst this was evaluated at monitoring sites (Figure 6.1), no monitoring attempt at individual level exposure evaluation was undertaken.

6.3.5 Representative errors

Exposure predictions for an individual or group, over a time period, can be made using dynamic models. But the degree to which these predictions are representative of that exposure

for that period of time is a source of uncertainty. Similarly, the degree to which predicting exposure for a group or groups of people, and extrapolating to the wider population, is a further source of confusion and error. In the LHEM, the exposures for each individual were calculated based on the persons previous days movements, and they were asked whether this was representative of their typical day. Analysing the responses to the latter question showed us that most of the subjects had a fairly set pattern of movement (and therefore likely exposure), but the degree of the variance between their typical day and a non-typical day is unknown, and indeed how many non-typical days do they have? One way to evaluate the exposure predictions from our model would have been to distribute personal monitors to all the LTDS participants, for an entire year, collect and process the data, and then compare it to the LHEM predictions to see whether our snap shot day exposure was close to their annual average day exposure. However, there were 45,000 subjects, meaning a huge number of personal monitors and logistical support would have been needed, and it was therefore unfeasible. A method to make this more manageable might be to develop a statistical model whereby an appropriate sample-size is calculated. That is, just sampling a percentage of the 45,000, but in the knowledge that they will have a similar distribution of exposures to the total population of the study; then giving this smaller subset of participants monitors for a prolonged period of time and understanding the daily changes in their exposure. This is explored more in Methods (Section 6.4)

6.4 Methods

We can now see that there are a number of difficult and challenging sources of error to consider in evaluating a hybrid exposure model. Air quality models have quantifiable errors in their predictions at the monitoring sites they are evaluated against, errors at locations away from monitoring sites which are not well understood, further variation and errors when using a high time resolution air quality model, errors in the micro-environmental modelling, and more errors in terms of the representativeness of the exposure predictions arrived at. The following methods section outlines how efforts were made to mitigate some of these sources of error, how some were removed altogether, and then produces an evaluation of one short journey by comparing the LHEM prediction of this journey to measured exposure using personal monitoring. The process is summarised below in Figure 6.3.

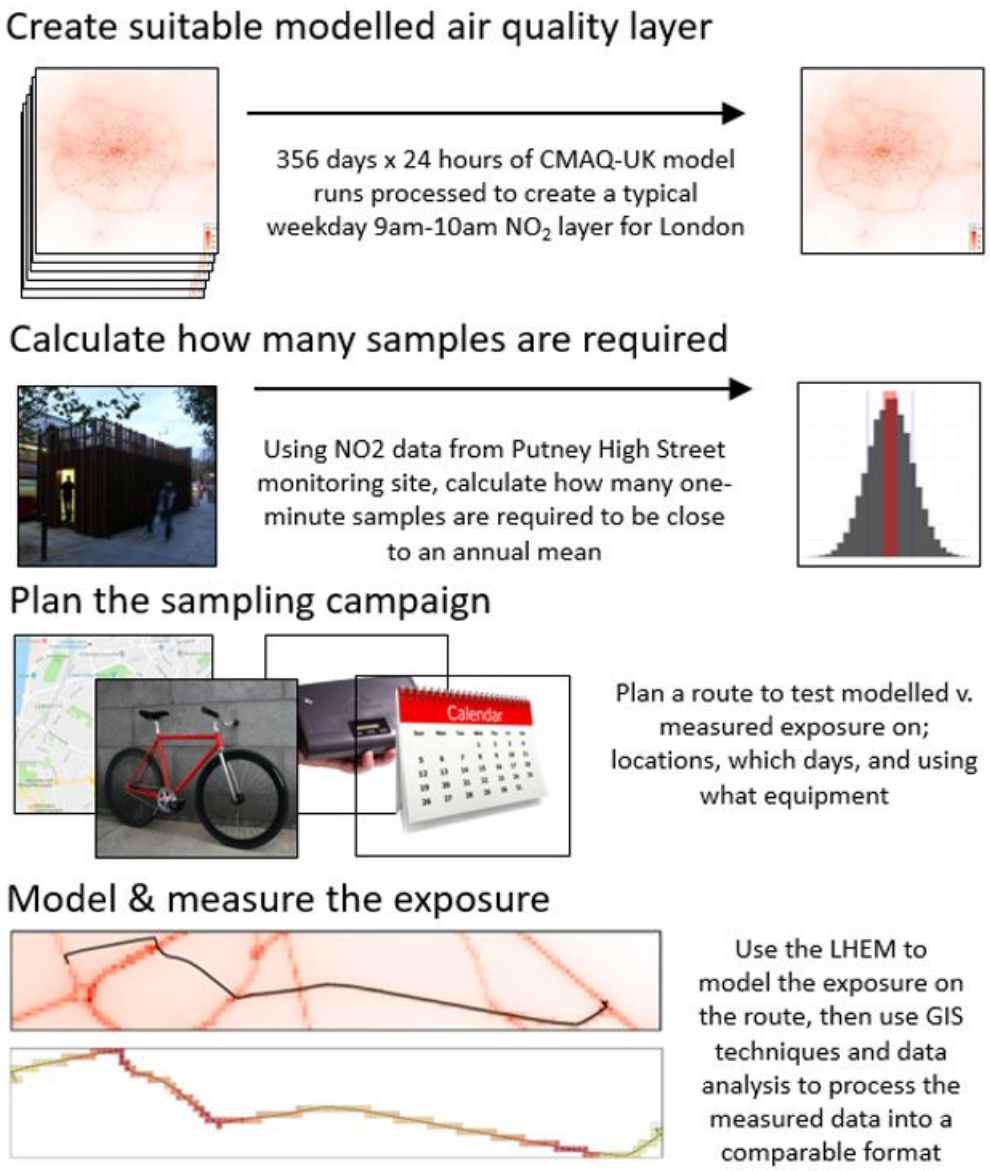


Figure 6.3: Method summary

6.4.1 Modelled Air Quality

The exposures in the LHEM contained Londoners movement data covering twelve months of the year between 2005 and 2011. The CMAQ-UK air quality model used was an annual average hourly weekday/Saturday/Sunday model of 2011 air quality, and linked to the movement of the individual based on what day of the week their time-activity data was recorded. The air quality model for this chapter is a 2016 CMAQ-UK output, with appropriate emissions and meteorology, cropped to London, but otherwise unchanged in methods from the 2011 model. When evaluated on an annual average basis i.e. one value for each 20m x 20m grid square of London, this new 2016 model performs well (Table 6.2).

Table 6.2: Annual average performance statistics for CMAQ-UK 2016

Pollutant	Number of data	Mean obs	Mean Mod	FAC2	MB	MGE	NMB	NMGE	RMSE	r	COE
CO	2	369.79	318.75	1	-51.04	72.08	-0.14	0.19	88.32	1	0.50
NO_2	73	51.80	48.49	1	-3.30	9.01	-0.06	0.17	13.79	0.84	0.52
NO_x	72	134.74	108.52	0.97	-26.23	39.52	-0.19	0.29	61.50	0.78	0.43
O_3	16	29.74	35.40	1	5.66	6.59	0.19	0.22	7.81	0.83	0.07
PM_{10}	56	22.76	22.74	1	-0.03	2.93	0.00	0.13	4.14	0.57	0.22
$PM_{2.5}$	23	13.08	12.34	1	-0.74	1.85	-0.06	0.14	2.30	0.72	0.20

We can see that the RMSE values for the pollutants range from 88.32 PPB for CO down to 2.3 μg m^{-3} for PM_{10}. For NO_2, which is the pollutant being focused on in this chapter, the RMSE is 13.79 μg m^{-3}. An exposure model using annual average NO_2 predictions from this 2016 CMAQ-UK air quality model could incorporate these errors; a prediction of 40 μg m^{-3}, presuming a normal distribution of errors, would therefore actually be anywhere within the range of 26.21 μg m^{-3} to 53.79 μg m^{-3}. The exposure-health relationships investigated by epidemiologists could then take this into account. Moving to a higher temporal scale, the daily-hourly version of this same model (Table 6.3) performs less well for NO_2 than the annual average shown above (Table 6.2).

Table 6.3: Daily-hourly performance statistics for CMAQ-UK 2016

Pollutant	Number of data	Mean obs	Mean Mod	FAC2	MB	MGE	NMB	NMGE	RMSE	r	COE
CO	24520	408.92	338.26	0.78	-70.65	175.22	-0.17	0.43	259.18	0.53	0.23
NO_2	656419	52.10	49.01	0.84	-3.09	18.63	-0.06	0.36	28.32	0.69	0.35
NO_x	649220	136.27	110.28	0.68	-25.99	70.72	-0.19	0.52	129.38	0.63	0.36
O_3	148043	28.84	34.31	0.66	5.46	12.88	0.19	0.45	17.20	0.74	0.30
PM_{10}	49461	22.96	22.87	0.89	-0.09	7.07	0.00	0.31	11.89	0.70	0.35
$PM_{2.5}$	188237	13.15	12.44	0.85	-0.71	3.72	-0.05	0.28	5.76	0.86	0.51

The NO_2 predictions have an R value down from 0.84 to 0.68 and a RMSE up from 13 μg m^{-3} to 28.32. Using a higher temporal scale model has increased the uncertainty of the air quality concentration predictions to the model.

To evaluate the LHEM using this air quality as an input, we needed to decide what time period of exposure to try and replicate with measurements. Evaluating at a higher time resolution than annual average was preferable, but a very high time resolution would be difficult as this would mean needing to take a huge number of measurements. For example supposing we wanted to evaluate exposure on Saturdays between 11am and 2pm and how the LHEM modelled this (within the context of a journey); measurements would need to be taken continuously on the journey in question every Saturday throughout the year between 11am and 2pm to compare the measurements with the modelled air quality output (putting the spatial and micro-environmental factors aside for a moment). To simplify the evaluation, and make the experiment a reasonable undertaking, the timeframe that it would

be performed over was restricted, so rather than cover a whole year we considered August and September only, on weekdays, and between the hours of 9am of 10am. To create an appropriate air quality model for this time period, 365 days of CMAQ-UK layers were used (each with 24 hours). A loop was written in R to download the raw CMAQ-UK NetCDF model files, convert them to standard raster R files, combine the individual files for each hour of the two months in question, discard data that was not 9am to 10am, and then take the mean for each grid square. The concentrations were outputted in PPB, so were converted to μg m^{-3} using the EU standard of 1.9125. Figure 3 shows the result of this process i.e. the models predicted air quality for a typical weekday in August or September between 9am and 10am.

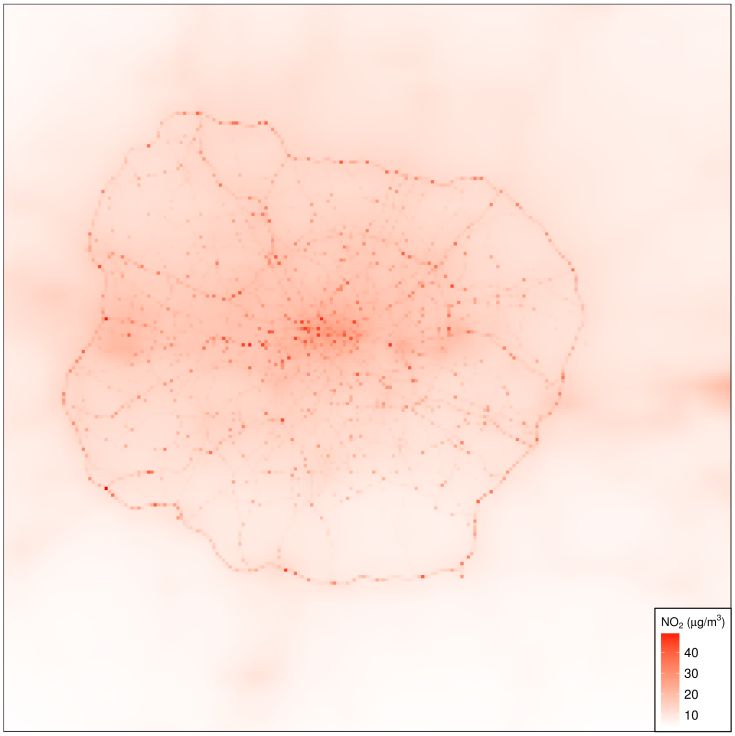


Figure 6.4: CMAQ-UK air quality model for 9am to 10am on weekdays in August/September

6.4.2 Micro-environmental adjustments

As discussed in the introduction to this chapter, and within Section 2.1.6, calculating exposure within micro-environments in relation to outdoor concentrations is not a well understood area of science and seems to be susceptible to a great deal of variation depending on conditions of the micro-environment (and other factors). To simplify the uncertaintity, this evaluation only considered the exposure of a journey which did not need micro-environmental modelling, and looked at journeys of the LHEM that take concentrations directly from CMAQ-UK i.e. walking and cycling. This is not to say that these areas an unimportant,

due to the large amount of time that people spend indoors they clearly are, but in an effort to compartmentalise this chapter in order to make the research achievable, these areas were removed.

6.4.3 Sample size

Having established that evaluation will only be undertaken for a cycling journey between 9am and 10am on weekdays in August and September, the number of cycling journeys (and measurements) now needed to be calculated in order to represent modelled exposure of this same journey. Repeat measurements were needed as although CMAQ-UK has well resolved temporal and spatial inputs such as vehicle numbers and emissions by time of day and day of the week, with similar meteorological inputs, it does not take account of inter-day variability of an unpredictable nature e.g. a car breaking down leading to a traffic jam, and increased emissions for an hour of a Saturday morning. By taking repeat measurements, the effects of these type of events on the evaluation should be reduced.

Considering this question in a theoretical framework, based around the LHEM, we could presume that the mean daily $PM_{2.5}$ exposures of the 45,000 LTDS subjects is 15 ug/m3, with a standard deviation of 2.5 ug/m3. We can then use a random number generator within these parameters to assign each subject an exposure, which creates a distribution shown in Figure 6.5.

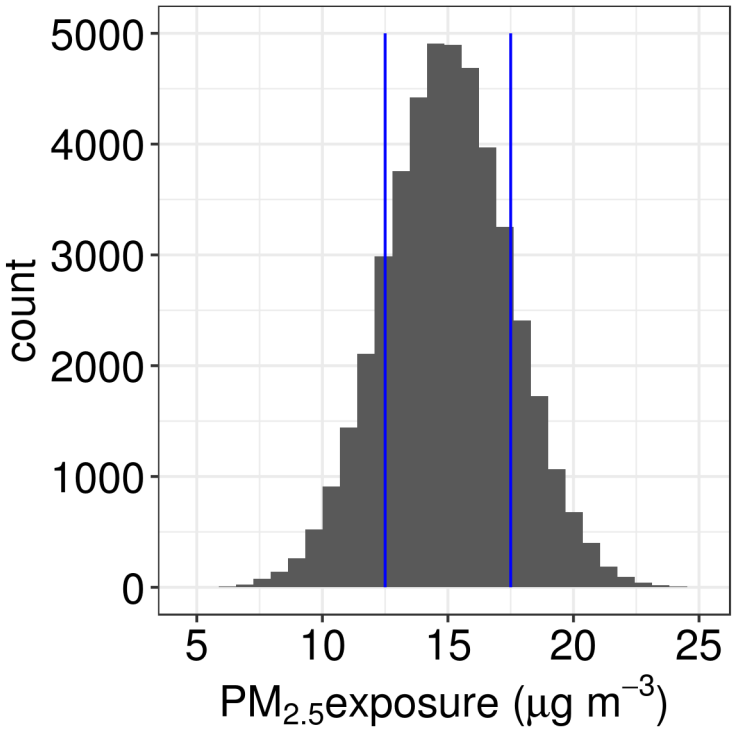


Figure 6.5: Theoretical LHEM exposures of 45,000 subjects based on a pre-defined mean of 15 μg m^{-3} and a standard deviation of 2.5 μg m^{-3} (shown in blue)

This distribution of exposures has a mean of 15 μg m^{-3}, a standard deviation of 2.5 μg m^{-3}, a max of 25.19 μg m^{-3} and a min of 5.61 μg m^{-3}. If these were true exposures of LHEM subjects, and measurements were required to evaluate a modelling attempt of these subjects, and the following parameters were deemed acceptable; a mean within 10% of the population mean (15 μg m^{-3}) with 95% confidence, then a sample size calculation (PennState Eberly College of Science (2017)) can be undertaken to arrive at the answer of 43. That is, only 43 subjects need to be measured to evaluate the model of the population in the theoretical exposure distribution. However this just gives a sample (n=43) distribution, that will have the same mean, that is within the area highlighted in red in Figure 6.6; but may or may not have the same shape of distribution, maximum, minimum, standard deviation etc. of the population. It is also likely to miss evaluating the exposures of the subjects who have exposures in the highest and lowest percentiles of the data, which as discussed in Section 2.2.2 may be important for understanding health effects.

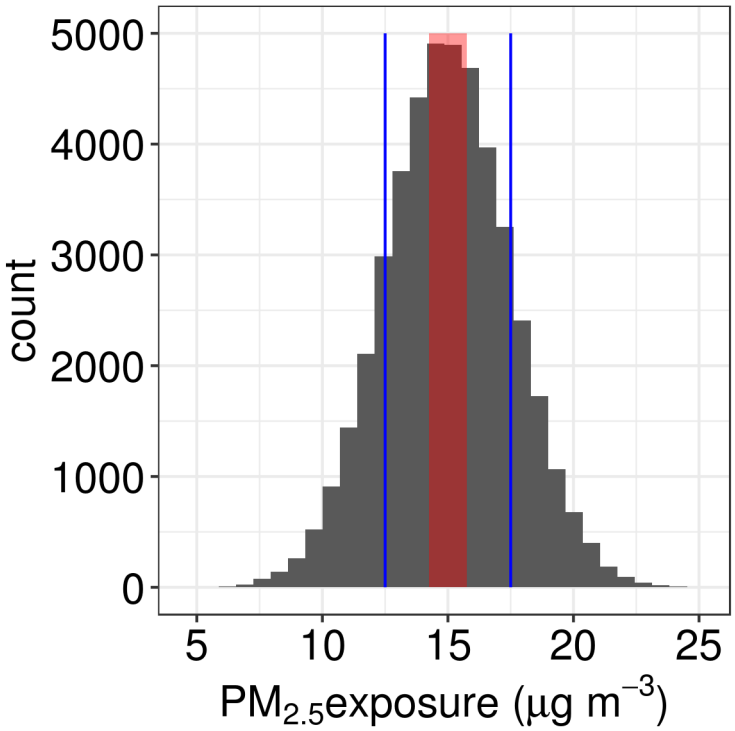


Figure 6.6: Theoretical LHEM exposures of 45,000 subjects based on a pre-defined mean of 15 μg m^{-3} and a standard deviation of 2.5 μg m^{-3} (shown in blue), with the mean of a sub-sample of 43 subjects (within the red area)

To properly capture the distribution of the population in a sample, stratified random sampling is required, to ensure proportional representation of each strata (exposure concentrations in this case). However to do this a priori understanding of the exposures of the larger group, and the drivers of the exposures, is required in order to be able to calculate the numbers of samples required from each strata.

Given these difficulties in trying to evaluate the typical day of an individual in the LHEM, and how this evaluation might be compared to an annual average exposure, it was decided to simplify and reduce the task by evaluating the modelled and measured exposure of a single cycling journey from the LHEM rather than a number of people or journeys.

Calculations to estimate how many measurements of the journey on a weekday between 9am and 10am in August and September would be needed to represent (within predefined boundaries) the typical exposure on the journey to compare to the model were undertaken. One-minute data from a monitoring site (as a known, continuous, and reliable dataset) was downloaded, and sample size calculations were undertaken, as a proxy for a CMAQ-UK grid square in the model. By taking one year of one minute NO_2 monitoring data from the site Wandsworth - Putney High Street, removing the data outside of the time period of interest (weekdays, 9am to 10am in August and September), and then calculating summary

statistics on the remaining data (Shown in Table 6.4). This site was chosen as the site collects one-minute resolution NO_2 data (the only one currently in London which does so), which aligns with the temporal resolution of the equipment used for the monitoring.

Table 6.4: Site WA7 summary statistics for weekday, 9am-10am during

Statistic	**NO_2 μg m^{-3}**
Minutes	2400.00
Mean	90.92
Median	77.70
Max	1074.40
Min	11.67
Standard deviation	64.23

Using these summary statistics for the period of time of interest, the number of samples required was calculated using Equation 6.1 (PennState Eberly College of Science (2017)) to create a similar distribution and mean, as per the theoretical example from Figures 6.5 and 6.6. By taking this reliable and continuous dataset, in a known location, the results were extrapolated to locations where there is not a fixed site data collection method available.

$$\chi = Z^2 \times \frac{\sigma^2}{moe^2} \tag{6.1}$$

Where χ is the calculated sample size, Z is the z-score of the desired confidence level, σ is the standard deviation of the population (concentrations), and *moe* is the allowed margin of error (in μg m^{-3}). By testing a variety of input variables to this equation, the number of samples required in different scenarios of confidence levels and allowable margins of error was calculated, shown in Table 6.5:

Table 6.5: Numbers of samples required for each confidence level and allowable margin of error based on measurements from WA7

Confidence level	Allowed margin of error	Sample required
99%	10% (± 9 μg m^{-3})	1325
	15% (± 14 μg m^{-3})	589
	20% (± 18 μg m^{-3})	331
	25% (± 23 μg m^{-3})	212
	30% (± 27 μg m^{-3})	147
	35% (± 32 μg m^{-3})	108
95%	10% (± 9 μg m^{-3})	767
	15% (± 14 μg m^{-3})	341
	20% (± 18 μg m^{-3})	192
	25% (± 23 μg m^{-3})	123
	30% (± 27 μg m^{-3})	85
	35% (± 32 μg m^{-3})	63
90%	10% (± 9 μg m^{-3})	540
	15% (± 14 μg m^{-3})	240
	20% (± 18 μg m^{-3})	135
	25% (± 23 μg m^{-3})	86
	30% (± 27 μg m^{-3})	60
	35% (± 32 μg m^{-3})	44
85%	10% (± 9 μg m^{-3})	414
	15% (± 14 μg m^{-3})	184
	20 % (± 18 μg m^{-3})	103
	25% (± 23 μg m^{-3})	66
	30% (± 27 μg m^{-3})	46
	35% (± 32 μg m^{-3})	34
80%	10% (± 9 μg m^{-3})	327
	15% (± 14 μg m^{-3})	145
	20% (± 18 μg m^{-3})	82
	25% (± 23 μg m^{-3})	52
	30% (± 27 μg m^{-3})	36
	35% (± 32 μg m^{-3})	27

This showed that in order to obtain a result which is within 10 μg m^{-3} of the mean, with 99% confidence, 1325 samples would need to be taken, at random, in the time window, in a grid square. Given the aim is to take measurements on a cycling journey, and as there are only 44 weekdays in August and September, attaining 1325 samples was not achievable. So it was decided to take 27 samples, which would give a 80% confidence interval (CI) and 32 μg m^{-3} margin of error (MOE). Clearly it would be better to have may more samples and stronger results with a higher CI and lower MOE, but this was not possible in the time-frame (and not essential as the focus of this research is to establish the methods and procedures that would be involved rather than to provide a comprehensive answer/evaluation). This

issue is explored further in the Discussion, (Section 6.6).

6.4.4 Measuring journey exposure

As collecting NO_2 data on 27 journeys at one-minute resolution would be difficult (Kings do not own a reliable and portable NO_2 sensor that can monitor at this frequency, indeed there are few commercially available at all). It was decided to collect black carbon (BC) data using an Aethlabs Microaeth AE51 (Hansen et al. (1984), Aethlabs (2016)), and to then convert these measurements to NO_2; the microaeth being a well understood and reliable device used in many exposure studies (Cheng and Lin (2013), Viana et al. (2015)). To calculate the conversion factor between BC and NO_2, NO_2 and BC data were downloaded from the Marylebone Road monitoring site (WA7) which collects both pollutants at 1-minute resolution, and then a linear regression model was created (Figure 6.7).

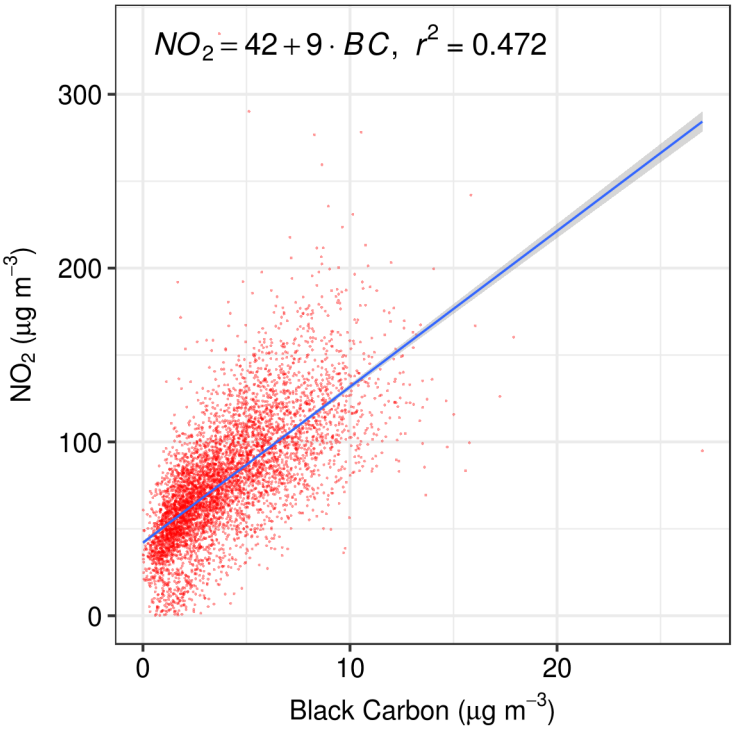


Figure 6.7: Linear regression between black carbon and NO_2

This gave a conversion of $NO_2 = 41 + 9.1$ x BC, with an adjusted R^2 of 0.474. A higher R^2 would have been preferable to give more confidence to the BC to NO_2 conversion process. This limitation is in the Discussion (Section 6.6).

With regards to further sources of error in the evaluation of the LHEM using monitored data, it was presumed that the Microaeth AE51 was giving perfectly accurate readings of

black carbon. Accepting that this is a limitation and in reality the device has been shown to have (relatively modest) measurement errors (Cheng and Lin (2013), Viana et al. (2015)).

A monitoring campaign was then completed over August and September, taking measurements on 28 cycling journeys between Kennington Park and Waterloo between 8am and 9am using the Microaeth AE51. The journey between Kennington Park and Waterloo was chosen due to convenience for the researcher and is purely arbitrary; the method could have been applied anywhere within the model domain.

6.4.5 Modelling journey exposure

Using the CMAQ-UK air quality layer as an input the exposure of the journey between Kenningtona and Waterloo was now modelled, using the same methods as used in the LHEM from Chapter 4. The route and concentrations are shown in Figure 6.8 below.

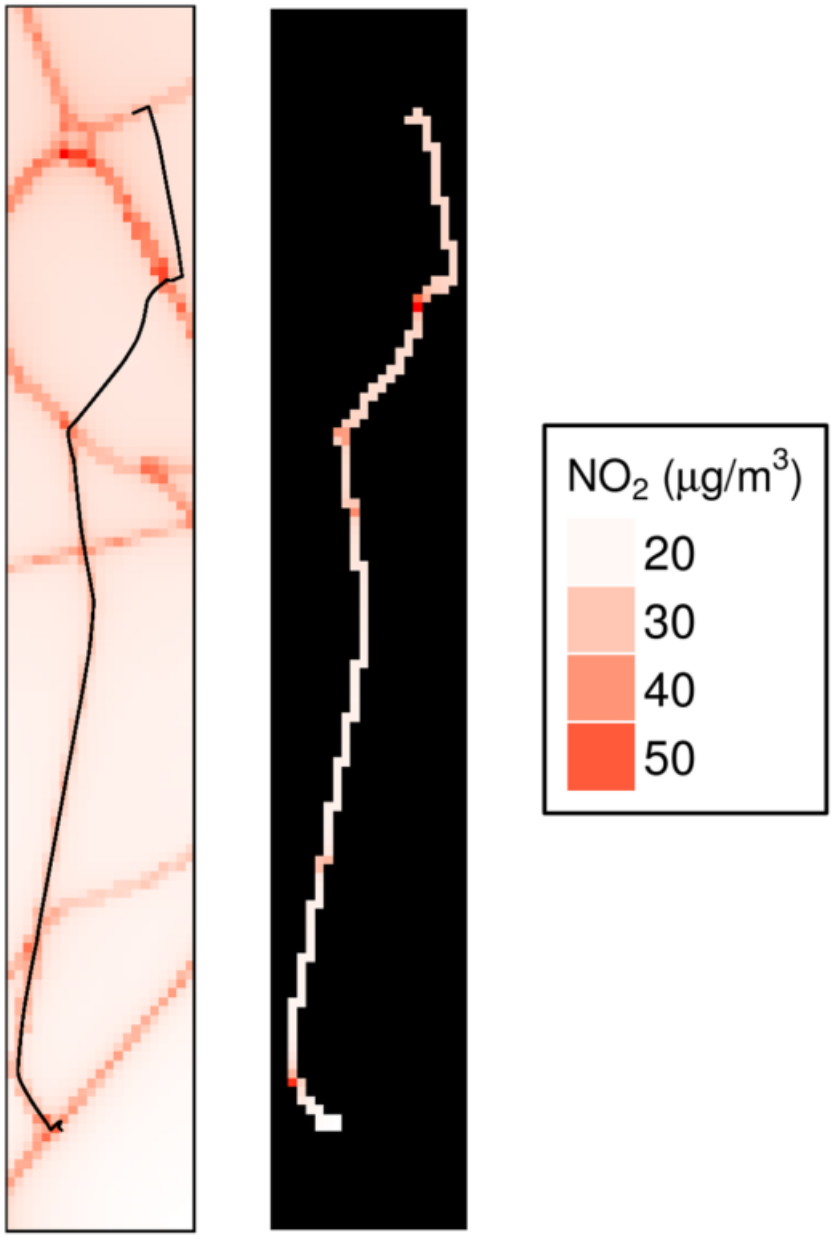


Figure 6.8: Cycling journey between Kennington Park and Waterloo, overlain on modelled CMAQ-UK concentrations for 9am to 10am on weekdays in August to September

6.4.6 Data processing

Before the measured exposure datasets could be compared to the modelled exposure of the same journey the following steps were completed:

- Raw CSV data from the Microaeth was downloaded, and unnecessary columns and header meta-data were removed.
- Black carbon concentrations were divided by 1000 to convert into micrograms per

metre cubed (for comparison with the NO_2)

- Missing GPS data was manually inferred and inserted (the device does not record position for first few minutes of operation, which was unknown at the time of data collection)
- Poor quality GPS data / drift was corrected by snapping points to the nearest road segment (Shown in Figure 6.9
- Line segments were created from the point positions.

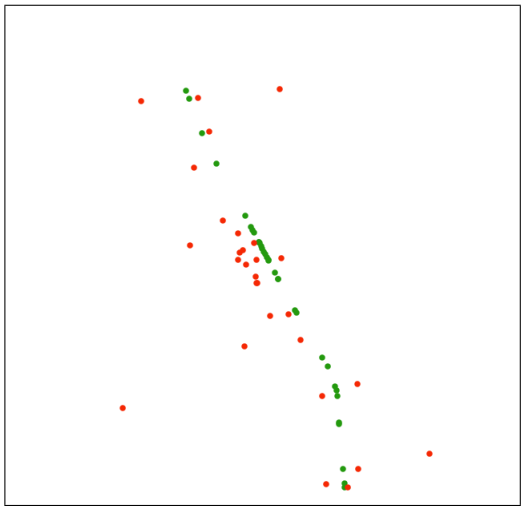


Figure 6.9: Locations of sampled BC concentrations along the cycle journey (red), which have been snapped to the road (green) to correct for GPS drift

The final step of creating line segments from points was required as the microaeth concentration are stored each minute, representing the end of one minute of sampling, meaning that the GPS position is the mean concentration from the previous minute of movement. To enable linking of these concentrations with the CMAQ-UK grid squares during that period of movement, a SpatialLinesDataFrame line between each GPS point, and the previous GPS point in that journey was created, and assigned the concentration of that whole line. The result of this processing of the monitoring data was 27 lines, stored as a SpatialLinesDataFrame (SLDF), split into segments of varying lengths, along the length of the journey from Kennington Park to the Waterloo. The start of one of the journeys is shown in Figure 6.10 below. Although not shown in the figure, the line has concentration attributes linked to it.

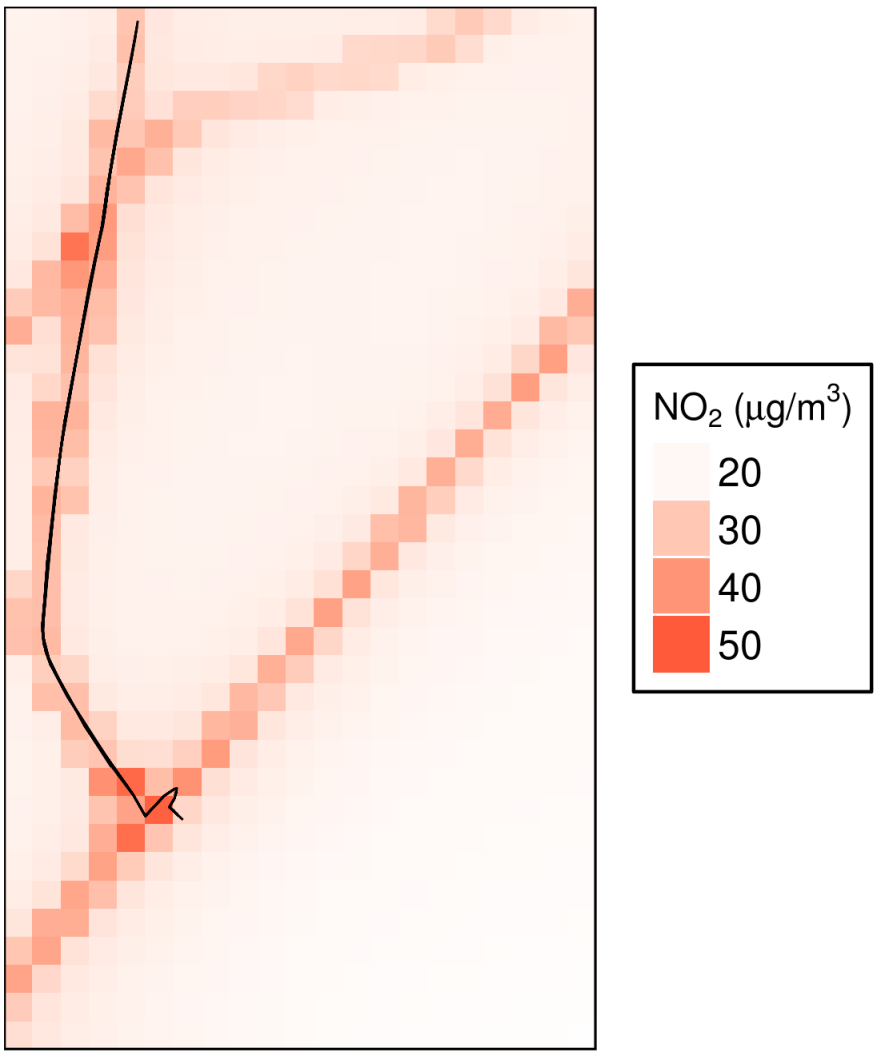


Figure 6.10: Start of a cycling journey from Kennington Park to Waterloo, cycle route shown by a black line, CMAQ model concentrations shown behind.

Using the SLDF of measured concentrations, and the CMAQ-UK raster of modelled concentrations, two result datasets were now created:

- **Concentration comparison**: The mean concentration from each journey, compared to the mean concentration of the grid cells that the journey intersected.
- **Spatial comparison**: For each grid cell the route intersected, the mean concentration from the 27 monitored journeys. Then, by grid square, the difference between monitoring concentrations and the modelled concentrations - output as a new raster file.

6.5 Results

6.5.1 Concentration comparisons

Figure 6.11 shows a boxplot summary graph of the 27 monitored journeys (right, black) compared to the modelled journey (left, red).

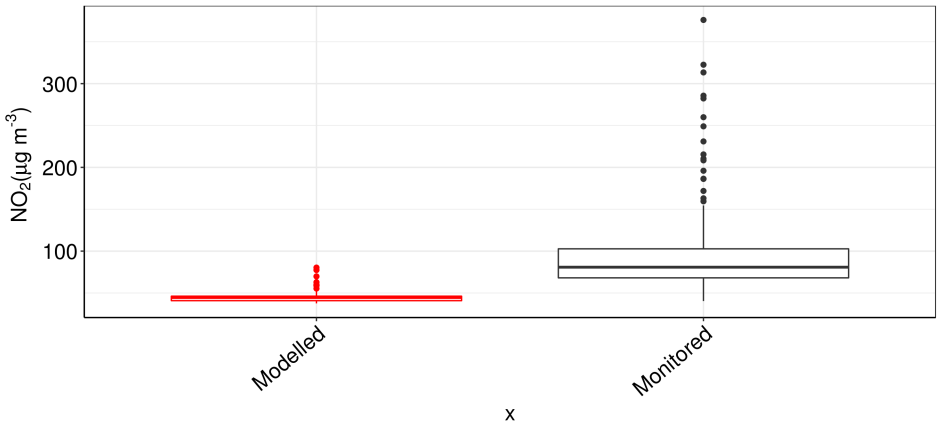


Figure 6.11: Box-plot of modelled cycling journey exposure compared to monitored cycling exposure

From this summary plot we can see that in general the monitoring campaigns found higher concentrations than modelling the journey. When visualised individually (Figure 6.12) we can see the variation more clearly. Noting that earlier we calculated that 27 repeat measurements were needed for a 80% confidence interval and 35 μg m^{-3} margin of error comparison, and should therefore really only be considering this data in aggregate form.

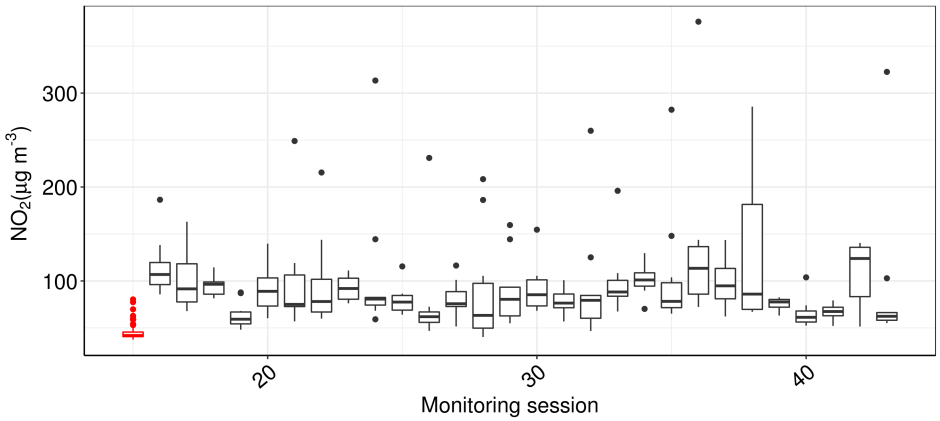


Figure 6.12: Box-plots of monitored cycling journeys (black) compared to the modelled cycling journey (red). Note that due to the device numbering system the sessions were numbered between 15 and 43, but there are not 43 sessions in the graph.

Note that due to the device numbering system the sessions were numbered between 15 and 43, but there are not 43 sessions in the graph.

The summary statistics for these two datasets are shown in Table 6.6. Taking the raw monitored data against the modelled data, the lower boundaries of the datasets are similar, but the monitoring found mean and median concentrations approximately double that of the modelling, and the maximum monitored value is way in excess of the model. If we consider the monitored data, and include the allowable margin of error defined in the sample size calculation of 30 μg m^{-3}, then the agreement of the means at the lower boundary of the margin of error are much closer (64.19 μg m^{-3} for monitored compared to 44.82 μg m^{-3} for modelled).

Table 6.6: Comparison of measured and modelled exposure on the cycling journey

	Modelled NO_2 μg m^{-3}	**Monitored NO_2 μg m^{-3} (MOE)**
Minimum	37.58	40.35
1st Quartile	40.48	68.08
Median	44.57	80.81
Mean	44.82	94.19 (64.19-124.20)
3rd Quartile	46.19	102.70
Maximum	80.33	376

6.5.2 Spatial Comparisons

For reference, figure 6.13 shows the journey from Kennington Park to Waterloo .

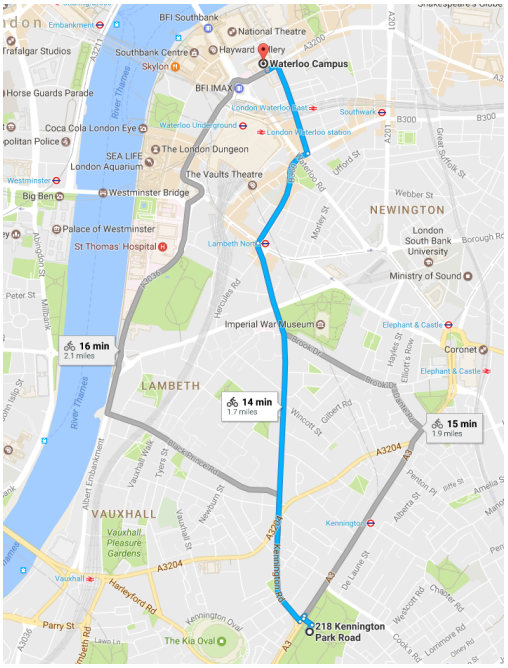


Figure 6.13: Map of cycling route between Kennington Park and Waterloo

Figure 6.14 below shows the monitoring data, the modelled data, and a comparison of the two, on the route.

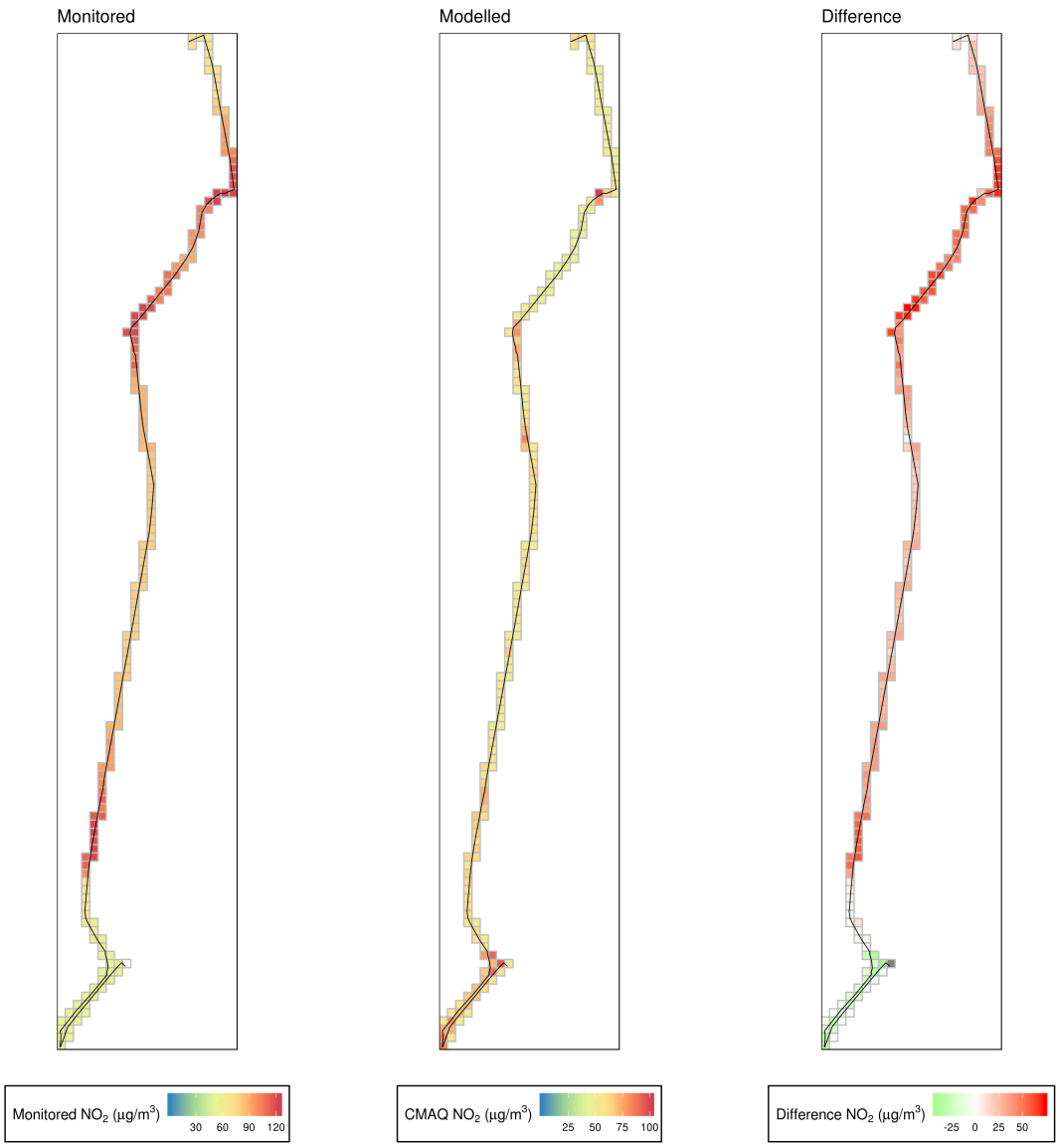


Figure 6.14: Map of monitored, modelled and difference concentrations along the journey (NO_2 derived from black carbon)

By spatially comparing the results in this manner (Figure 6.14) we can see that the monitoring campaign (left) found the highest concentrations (> 100 μg m^{-3}) around the junction of Kennington Road and Kennington Lane (towards the South of the journey), near Lambeth North Station (a busy junction), and again near The Old Vic theatre (another busy junction). These locations of high concentrations are also found in the modelled data (centre), but the difference against other stretches of the journey is not as large, and compared to the monitored data the concentrations are not as high. Considering the difference map (right), except for the first 5-10% of the journey, most of the journey is underestimated by the model, between 0 and 25 μg m^{-3} in each grid square. The largest differences are on Baylis Road, between Lambeth North Station and the Old Vic theatre, where the monitoring

found concentrations to be up to 50 μg m^{-3} higher than the model.

6.6 Discussion

As discussed in the Background (Section 6.3), to my knowledge, there have not been any studies which have attempted to evaluate a dynamic approach to exposure in this manner. Studies have tended to measure certain micro-environments for a time-span often determined by practical constraints such as battery life, staff time and convenience; and then use these as empirical comparisons to their model outputs.

This piece of research was novel in that exposure was modelled, and then an attempt to evaluate the predictions (of an example journey) was undertaken by calculating how many personal monitoring samples would be needed, and then going out to collect that data. Despite various sources of possible error, this research demonstrated this as a possible process, and highlighted the difficulties of it. The results were not particularly encouraging in terms of comparing modelled v. measured absolute values, but clear spatial patterns were discernible.

There were many issues and challenges to overcome whilst undertaking this evaluation that will have impacted the accuracy of the results. The main one being (mostly unavoidable) sources of error at each stage of the process. On the modelling side of the comparison CMAQ-UK has been shown to perform well when evaluated against monitoring sites, but it is far from perfect - the input the modelling uses for this chapter has NO_2 R values of 0.75. Against this, it was calculated that 27 samples would be needed to get a monitored concentration estimate that had a 30 μg m^{-3} MOE (and 80% CI). Black carbon samples were collected to do this, using a MA300 Microaeth, which has not yet been evaluated by academic literature (although previous models of similar equipment has R^2 values of 0.8), and these data were converted to NO_2 using a linear regression with an R^2 of 0.47. These multiple sources of error are bound to have confounded the results, but to what degree is uncertain. Post-hoc, it is not possible to go back and improve the CMAQ-UK model, or to collect more samples, or to improve the accuracy of the portable monitor, but it is interesting to reconsider the conversion process of black carbon to NO_2 from Figure 6.7. Visually the intercept of 41 looks high, and if this was theoretically changed to 0, then the measured concentrations would all be reduced, and the comparison between modelled and monitored would be much closer. Figure 6.15 below shows a revised box-plot comparisons in this situation.

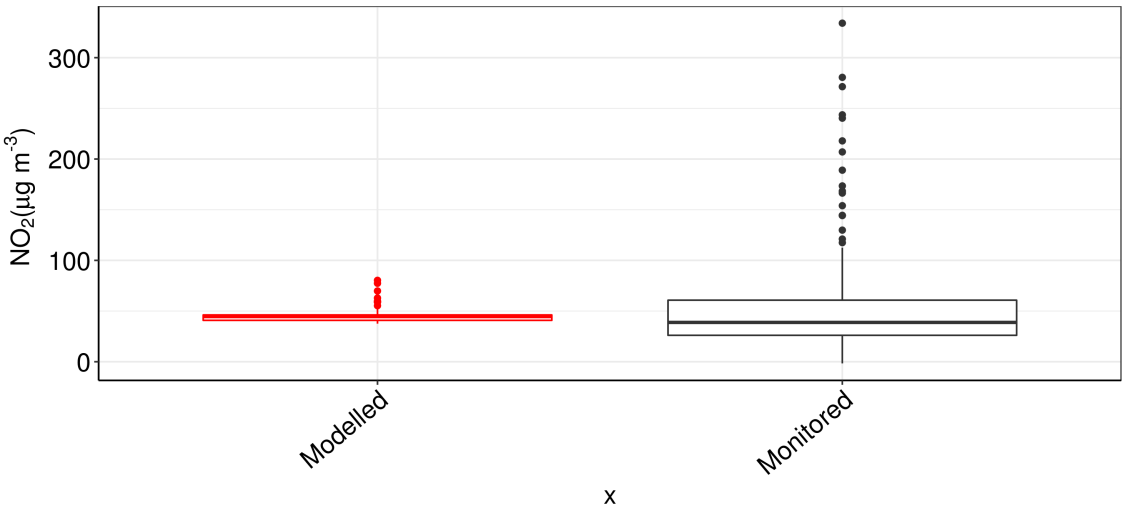


Figure 6.15: Boxplot comparison of modelled and monitored concentrations with an intercept of 0 between BC and NO_2

Putting measurement errors and modelling errors aside, there are practical improvements that could be implemented in repeating this work which might lead to improved results. The most notable being the sample size. If the journey had been shorter, or more researchers were available, it would have been possible to collect a much larger number of samples, which according to the sample size calculations would have increased the reliability of the results. A more practical long-term approach to evaluating air quality model inputs to a hybrid exposure model could be to fix reliable portable monitors to cars or public transport and increase sample size in this manner. For example, mobile monitors might be attached to the outside of a fleet of No.59 buses which run along the route monitored, meaning each grid square would be sampled 10-15 times between 9am and 10am, resulting in 400-500 measurements over the same time period; within King's colleagues are already trialling a similar method to this, but with the monitors inside a bus to measure the changes in passenger exposure from the electrification of a bus route in London.

Focusing on the equipment, an unforeseen issue in collecting the monitoring data was a mismatch between cycling speed, the resolution of the CMAQ-UK grid squares, and the temporal resolution of the Microaeth MA300. The aim was to measure concentrations in each grid square along a route, but as the Microaeth was set to 1-minute resolution, the grid squares were 20m by 20m, and typical cycling speeds are 15 km/h (or 250 metres per minute), the result was that each minute of Microaeth data often covered multiple grid squares. An improvement would be to calculate the sampling rate of the device based upon the likely speed of the movement, and the air quality model resolution it will be compared

to, as per equation 6.2

$$samplingrate_{(minutes)} = \frac{AQ_{(metres)}}{speed_{(metres/minute)}} \tag{6.2}$$

Though this would of course be constrained by the settings available in the sampling device.

Within the GIS and data processing section of this work, the main issue was the lack of accuracy of the GPS points that were output from the device, how to snap these appropriately to roads (presuming an available roads dataset), and how to link this to CMAQ-UK concentration grid squares for analysis. This was time-consuming and required manual editing at times. Automating this for a large scale campaign would be challenging but necessary. Junctions were a particular difficulty, for instance snapping GPS points at a left-turn always means that the point is snapped to part of the road either side of the junction and never at the actual junction itself (Figure 6.16); which might be where the highest concentrations occur and the researcher is most interested.

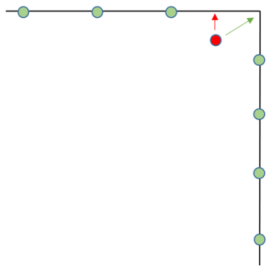


Figure 6.16: Accurate GPS points shown in green, GPS error shown in red. Correction should move the point to location shown by green arrow, but simple snapping will move it to direction of red arrow.

6.7 Conclusions

The main conclusions from this chapter are method-orientated, rather than definitive answers about the reliability of a dynamic exposure model as demonstrated in the previous chapters. The amount of data collection and data processing were both extremely onerous in order to even evaluate one journey, requiring skills in data processing,spatial analysis, statistics and personal monitoring. To repeat this method on a larger scale initial consideration should be given to the monitors that are going to be used, and in which possible transport modes, to ensure that they provide a high enough temporal and spatial resolution for alignment with the modelled air quality being used by the exposure modelling.

The sample size calculations here demonstrated a method (given a fixed reliable base dataset) to estimate the number of samples that are required to arrive at monitored concentrations that have a low margin of error and high confidence. For this piece of research, it was calculated that over 1,000 measurements would have been needed to give a low margin of error with high confidence, but it is worth noting that this figure depends on the pollutant of interest, the base data, and the geographical area.

Regarding results rather than method, the CMAQ-UK NO_2 model used as an input to the London Hybrid Exposure Model appeared to underestimate concentrations for the journey examined. Whether this can be extrapolated to draw wider conclusions about the model are unclear; the findings have large caveats due to multiple, mostly unavoidable, sources of error in the method (e.g. the black carbon to NO_2 conversion and sample size) that require fuller exploration of their impact and propagation.

7. Discussion, conclusions & future work

The original hypothesis of this research was that "*Individual and population level air pollutant exposure can be estimated using time-activity surveys, GIS and routing tools, and then coupled with high resolution spatio-temporal air quality models to facilitate a greater understanding of the health impacts of air pollution and how public health risks can be reduced*". This was expanded with the following objectives:

1. Create a model of Londoners daily movements based upon freely available TfL datasets
2. Link modelled air quality, estimate exposure, and compare with traditional methods
3. Create an exposure model for exposure to $PM_{2.5}$ on the London Underground
4. Develop an understanding of methods to evaluate predictions of exposure from hybrid-type models

This research has gone a long way to proving this hypothesis through the creation of a tool that allows scientists to examine individual-level and group exposure to ranges of pollutants in a range of microenvironments. The chapters on 'Modelling Londoners movements' (Chapter 3) and 'Dynamic exposure modelling' (Chapter 4), which were the main chapters centred around the creation of the tool, have been published as Smith et al. (2016) (the model development and use), and then subsequently used by Tonne et al. (2018) to examine socioeconomic and ethnic inequalities in exposure to poor air quality.

The chapter 'Exposure to $PM_{2.5}$ on the London Underground' (Chapter 5), on refining the exposure estimates for users of the London Underground after this micro-environment was found to be so important to Londoners exposure, is in the process of being written-up as an academic paper and will provide an open-source dataset ('TubeAir') for calculating the exposure of London Underground users within other studies, or as a stand-alone policy tool (probably for use by TfL and their contractors).

Finally the chapter on 'Evaluating dynamic exposure models' (Chapter 6), on how to evaluate exposure models of this kind, diverted away from the main hypothesis of the research, but during the previous chapters the reliability and how representative the results are became of interest and importance to consider, and so this seemed necessary. This final chapter concluded by finding that depending on the level of confidence and margin of error, large

numbers of personal monitoring data are required to evaluate dynamic models.

7.1 Discussion

7.1.1 The LTDS-X

The chapter on 'Evaluating dynamic exposure models' demonstrated how a dataset that was originally purposed for assessing transport demand in London could be adapted, processed and re-purposed to create the LTDS-X, a high resolution spatial and temporal time-activity dataset (including demographics), that is representative of the daily movements of the population of London. The main limitation with using this dataset as an input to the exposure modelling was that the data were London-centric, and 'porting' this method to another city or country would require a similar or replacement dataset. Though as the model is mechanistic rather than empirical, theoretically this should be possible. Whilst undertaking this research other datasets have been investigated and considered as substitutes for the LTDS, for example in the 2011 UK Census, a new question of 'workplace zones' was asked of the population, and research by Dr Reis (Centre for Ecology and Hydrology, UK) is using the responses as a way to model population-level movement for exposure modelling. Using this as a basis, travel exposure from workplace zone to workplace zone by each mode of transport could be simulated and the exposure calculated, and then an indoor exposure module 'plugged-in' to create a UK-wide exposure model. The spatial and temporal detail would not be as high as the LTDS-X and LHEM, but the coverage and number of subjects would be much larger. Given Census' are common in many countries around the world, this could be replicated in other places given time and processing effort. Other datasets that might be useful in simulating subjects movements include location data from such sources as storecards, credit cards, smart travel cards e.g. Oyster cards, geo-referenced tweets and phone signal data. Though balancing quantity of data against quality will be an issue in using many of these data sources. Nyhan et al. (2016) for example used mobile phone signal triangulation in New York to create an exposure model for 8.5m people, however the spatial and temporal resolution was very coarse. In obtaining and using this type of data researchers must also be aware of the laws and regulations which govern its use; and only use/share the data in ways that have been permitted by the original individuals.

It is worth noting that whilst the LTDS dataset was specifically re-purposed for use as an input to air quality exposure modelling, it could similarly be used for exposure to other things such as perhaps ultra-violet light and risks of skin-cancer, or to estimate where the population are drinking water and therefore links between poor quality water and health. Any type of model that requires human time-space-activity as an input.

Away from exposure the dataset has been useful in areas of work that the Environmental

Research Group is involved in. In 2012 it was used to look at the number of cross-Borough car trips being undertaken as part of the London Atmospheric Emissions Inventory (LAEI), and in 2018 was used to identify whether active travel has increased or decreased in the London Borough of Waltham Forest as part of a contract piece of work.

7.1.2 Dynamic exposure

In the chapter on 'Dynamic exposure modelling' (Chapter 4), having built the LTDS-X, this was combined with CMAQ-UK and microenvironmental modelling methods to quantify at unprecedented detail the exposure of the London population to poor air quality (The LHEM). The size, detail and possibilities for future research of this model were demonstrated by focusing on exposure missclassification, and this work was published in Smith et al. (2016). The applications for this model are already being taken forward in other exposure studies, including the project 'CLUE II' led by Imperial College London looking at the air quality and noise exposure of children in London, and the COPE (Characterisation of COPD Exacerbations using Environmental Exposure Modelling) study led by KCL. As was highlighted in the Chapter, it is our hope that this type of tool can increasingly be used for policy applications by the likes of TfL and the Greater London Authority (GLA) to better understand the effects of policy interventions on exposure. Indeed, Waltham Forest are using it to investigate the effects on cyclist exposure of introducing segregated cycle lanes, and we are collecting data on behalf of TfL to quantify changes in bus passenger exposure following retrofitting of cleaner engines.

The areas of the LHEM model that were identified as most in need of improvement were the modelling within microenvironments, which included time in enclosed transport (bus, car, train, tube), and time indoors. There are a number of studies that consider these microenvironments, but the results are highly variable. Similarly, little was known about the exposure of passengers on the London Underground while making the LHEM. When the initial LHEM results were being analysed it transpired that this was an important determinant of high daily exposures in the population, and therefore the objectives/research plan for the following chapter were developed.

7.1.3 The London Underground

The chapter 'Exposure to $PM_{2.5}$ on the London Underground' (Chapter 5) involved an extensive mobile monitoring campaign on the London Underground, and creation of a dataset

for estimating $PM_{2.5}$ passenger exposure during journeys on the network. Other research has been undertaken in the London Underground, but not to this spatial coverage, not geographically referenced, and not made publicly available. That is not to say it cannot be improved; there are a small number of stations that are not yet mapped, and specifically stations with multiple lines need further sampling and data refinement to separate these into different exposure estimates. Also further sampling could be undertaken to establish the repeatability of concentrations. This research has directly led to a sub-group of COMEAP being formed, and a report on "available evidence on the health risks associated with particulate matter exposure in the London Underground" being written, aswell as TfL commissioning further monitoring on station platforms.

Within the Environmental Research Group at KCL, Dr Green has also been contracted by TfL to undertake further monitoring on station platforms. The data and research should be published later in 2018 which is expected to lead to further opportunities and incorporation into other studies. For example the exposure tool was used to create an origin-destination matrix of exposure between stations on the Northern Line of the London Underground, shown in Figure 7.1 below .

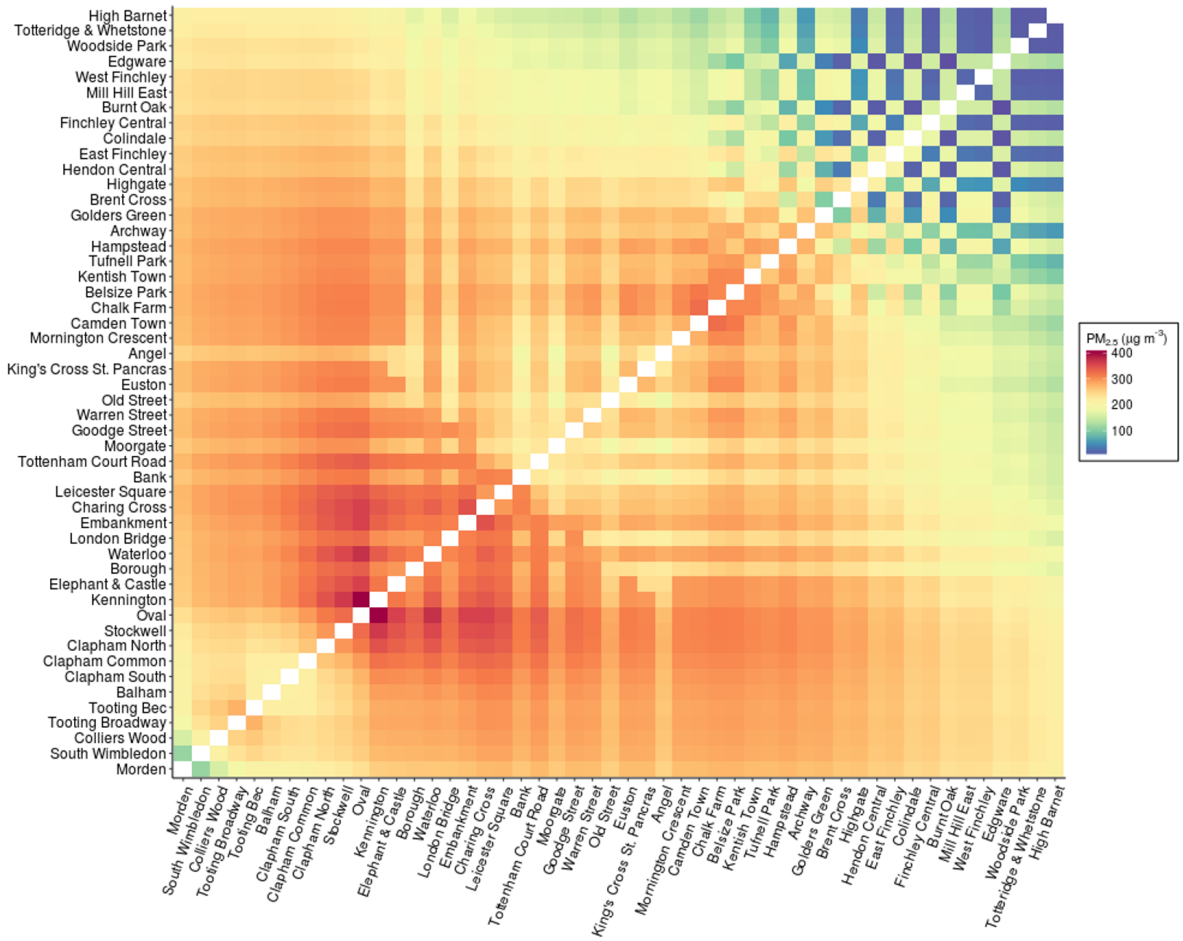


Figure 7.1: Time-weighted exposure between tube stations on the Northern Line

7.1.4 Evaluating dynamic exposure models

The area of research in the chapter on 'Evaluating dynamic exposure models', has been poorly addressed in publications in this area. It is common for authors to comment on the accuracy of their outdoor air quality model, how representative the microenvironmental modelling is, the quality of the time-activity data, or other such factors. But not to consider how to combine these factors, and not how to undertake sampling of journeys for comparison. The chapter on 'Evaluating dynamic exposure models' (Chapter 6) doesn't solve this problem, but demonstrates the issues and takes a first step at addressing one of them; how many samples of air quality are needed to be representative of annual averages across a journey (and the issues with collecting this data). It lends weight to the 'distributed sensor network' notion that has been mooted in publications such as Moltchanov et al. (2015) and Broday et al. (2017), and how these could be useful in exposure science rather than simply as a new type of monitoring site. By taking continuous mobile measurements all over a city, for example on all buses and bike-hire schemes, these repeat samples could be aggregated and a better understanding of hyper-local air quality developed to evaluate air quality modelling (and exposure).

7.2 Conclusions

7.2.1 Modelling Londoners movements

- While there are clear peaks in the morning and evening, there is substantial travel in London outside of peak hours
- Men tend to travel longer distances than women each day
- As household income increases, daily travel distance increases
- Travel distance increases with age, peaking around 40, then declines.
- The very old ($>$80) tend to only take short journeys over small distances.
- People over 80 rarely use the London Underground
- Between 20 and 50 years old, people spend about 70% of their time within 1km of their home.

7.2.2 Dynamic exposure modelling

- Londoners are exposed to 85% of their daily NO_2 and 90% of their daily $PM_{2.5}$ while indoors
- Increasing active travel results in daily lower exposure to both pollutants.
- Epi studies are likely overestimating exposure when using address point concentrations.
- There seems to be little difference between postcode and address-point exposure estimates
- Exposure missclassification increases in relation to the amount of travel that people undertake, especially inactive travel
- Londoners are exposed to peaks of unacceptable (according to the WHO) $PM_{2.5}$ and NO_2 levels for between 13-16% and 1-4% of their day, depending on age group.
- $PM_{2.5}$ and NO_2 exposure is correlated at the address level, but not in a dynamic exposure model.

7.2.3 Exposure to $PM_{2.5}$ on the London Underground

- $PM_{2.5}$ concentrations on the London Underground vary between 0 μg m^{-3} and 990 μg m^{-3}, with a mean of 129 μg m^{-3}.
- There are large variations between lines, and between sections of track on the same line.
- When ranked in decreasing order of mean $PM_{2.5}$ concentrations, London Underground lines have the following concentrations:

Victoria	436 μg m^{-3}
Northern	219 μg m^{-3}
Piccadilly	199 μg m^{-3}
Bakerloo	164 μg m^{-3}
Central	119 μg m^{-3}
Jubilee	115 μg m^{-3}
Metropolitan	67 μg m^{-3}
Circle	41 μg m^{-3}
District	40 μg m^{-3}
Hammersmith & City	36 μg m^{-3}
DLR	18 μg m^{-3}

7.2.4 Evaluating dynamic exposure models

- Evaluating dynamic exposure models requires expertise in personal monitoring, uncertainty analysis, statistics, air quality modelling and geographic information sciences.
- For 20m by 20m modelled air quality representing one hour in August and September, it was estimated around 540 measurements would be required in each grid square to give a 90% confidence with 10% margin of error.
- As temporal resolution of air quality models increase, the number of samples to evaluate the data increases.
- CMAQ-UK appears to be under estimating air pollutant concentrations (and therefore exposure) on the route chosen for evaluation

7.3 Future work

The natural direction of dynamic exposure modelling of this kind is to expand the geographical coverage, increase the numbers of subjects, the accuracy of the modelling (both in terms of the exposure predictions and how representative of populations the results are), and to then apply the model to human health studies. There are also a number of methodological ways that the model could be improved which will enable the model to be used quicker and more effectively. These concepts are explored below.

7.3.1 Routing improvements

The creation of the LTDS-X in Chapter 3 was mostly a result of cleaning of survey data, loading into database software, and then using the PL/R PostgreSQL language to communicate between the database and the routing API's through an R interface (with some further data cleaning and processing after this). Whilst this worked, it took a long time to run due to queries failing, and code needing to be re-written to anticipate and deal with various errors elegantly. Other issues included limits on API use meaning queries had to be submitted with pauses inbetween them to purposefully slow down the requests. Going forward this should be turned into stand-alone R code, that queries a range of APIs depending on the rate limits and transport mode, and ideally queries it's own routing server (such as OpenStreetMap (OSM) where no limits would be enforced - the Environmental Research Group are currently looking into setting up an ERG-OSM server. Not specifically for use by a hybrid exposure model, but as a general all-purpose tool for geographical datasets. Some of the routing code used in the LHEM was turned into a stand-alone routing/exposure tool for individual journeys. An R function was created whereby London start and end coordinates could be entered, and a PDF output of exposure on that journey created (Example shown in Figure 7.2).

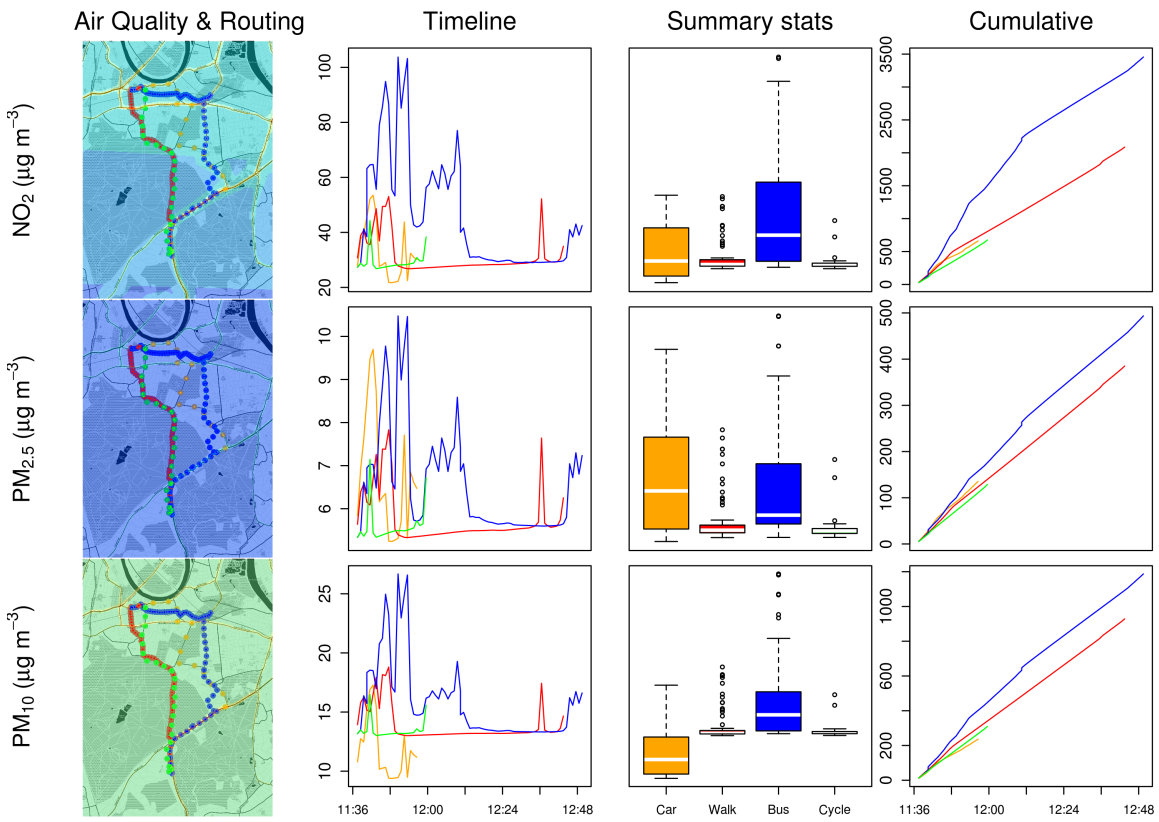


Figure 7.2: Example of stand-alone routing tool developed for Drayson

Many of the routing functions that have been created could be considered for adding to a stand-alone R library or package, either newly developed or to an existing package such as 'dodgr' (Godfrey and Govindaraju (2018)) or 'stplanr' (Lovelace and Ellison (2018)).

7.3.2 Geographical and temporal coverage

A logical step for this model is to expand it's geographical coverage, and to incorporate larger numbers of subjects. The LTDS survey used in this research covered the years 2008 to 2010. However there is now data available from TfL up to 2016. In addition, TfL have re-calculated the way in which they use scaling factors to represent the population, meaning that longitudinal analysis between years is now possible, such as calculating how cyclists in London's exposure has improved between 2008 and 2016 (providing air quality data for each year is available). Outside of London, using UK census data to increase subject coverage would also be desirable, perhaps as a complimentary dataset to the LTDS, rather than as a replacement. By combining the two, modelling of exposure would be detailed in cities where higher resolution datasets are available, and less detailed where it is not, in a similar way to some of the CMAQ-UK air quality modelling.

7.3.3 Application in health studies

This model was designed as a tool to better estimate exposure to poor air quality on the population of London, to then be able to take these findings forward into estimating the effects that this exposure may or may not be having. Health studies such as the one commissioned by the GLA in 2015 to look at the effects of NO_2 and $PM_{2.5}$ on Londoners (Walton et al. (2015)) was based upon modelling air quality concentrations at the Output Area (OA) level, combining this with population data, and then using these data as inputs to a health impact tool. Using the LHEM, adjustment factors at the Output Area level could be calculated, and these findings re-evaluated. Work is currently underway with a colleague to do this for a cohort of elderly Londoners in the English Longitudinal Study of Ageing (`https://www.elsa-project.ac.uk/`). The LHEM data is (in simple terms) being used to create a regression model between exposure and percentages of time in each micro-environment that the elderly inhabitants of each geographical area of London live in. This will then be expanded to a large cohort for which this contextual information is available.

In general however, methods to take this model forward into health studies needs more consideration and development. It may be more appropriate to re-design epidemiological studies afresh to incorporate this type of model from the start, rather than trying to apply it to already collected health data i.e. moving away from calculating exposure by geographical areas (postcode, output area etc.) to be individual based. At this level, the LHEM can provide much richer sources of exposure data than traditional models, for example by being able to enable consideration of short-term exposure at much shorter periods than days (days being 'short-term' in other studies such as Kloog et al. (2012) and Beverland et al. (2012a)). Indeed as part of the CLUE II project led by Imperial College London we are going to attempt to use the LHEM to predict the exposure of 150 school-children based on a dairy of their previous 24 hours activity, and then consider these results in the context of biological samples being collected. This has potentially exciting findings; perhaps we can see clear differences in the children's samples depending on who used the London Underground that morning compared to who cycled, and then be able to extrapolate these findings to the rest of the school.

7.3.4 Reproducibility

As the tool was mostly written inSQL (PostgreSQL + PostGIS), and the data stored in tables on an internal department server, the methods and data within it require a level of expertise and familiarity to work with. A new researcher could in theory recreate the

tool given available data, but it would be difficult. Editing the parameters of the model, for example the indoor to outdoor infiltration rates, and then re-running for new results, also require in-depth knowledge of the structure of the database. In retrospect, now with additional skills in languages such as R, RMarkdown, Latex and Python, and with a better understanding of the advantages of reproducible research, it would be hugely beneficial for the exposure tool to be re-built from the base up within a more flexible framework and language. Doing so would undoubtedly lead to speed improvements, allow for easier 'tweaking' of inputs i.e. a different years air quality, and more produce more concise analysis and outputs. It could also be developed within GitHub to allow contributions from other members of the department in a version-controlled manner. Following redevelopment, an interactive webpage to allow interrogation of the results would then ideally be built. Likely using the Shiny package (Chang et al. (2018)) within R.

7.3.5 Scale

As discussed in the introduction, most air contemporary quality exposure models tend to be at the postcode or address point level, with no dynamic aspect to them, typically annual average air quality estimates, and no micro-environmental modelling. One of the themes of this research was that by more accurately (spatially and temporally) predicting exposure, at the individual level, it would become possible (going forward) for epidemiological studies to better understand the health effects of poor air quality. This would theoretically be undertaken by linking individual health records to exposure estimates for large groups of people, and then some form of cohort or time-series analysis undertaken. It is worth note, that whether this will actually lead to a better understanding of the health effects of air quality is not a given. It might transpire that even once these high resolution air quality models, linked to health models, are created, the results may still be very unclear. Or find exactly the same relationships that were estimated 20 years ago. Postcode, address-point, Borough-level or city wide exposure estimates may be found to be unimportant, and perhaps it is just the amount of time that someone spends in a car which is the strongest determinant of exposure? It will be interesting to see how this field develops.

A joint study by King's College London and Imperial College London is currently being undertaken which may shed some light on this (Moore et al. (2016)). The study has only just begun, but Figure 7.3 below summarises how the research is planned to proceed.

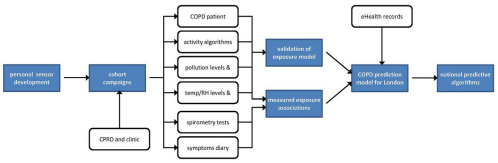


Figure 7.3: Method diagram for the COPE study

In brief, as part of the 'characterisation of COPD exacerbations using environmental exposure modelling' (COPE) project the methods of this thesis are being used to estimate exposure of a number of patients. These same patients will then have their exposure monitored (to validate the modelling), and then their symptoms monitored alongside their exposure, with links e-health records.

Bibliography

Adams, H., Nieuwenhuijsen, M., Colvile, R., McMullen, M., Khandelwal, P., 2001a. Fine particle (PM2.5) personal exposure levels in transport microenvironments, London, UK. Science of The Total Environment 279, 29–44.

Adams, H.S., Nieuwenhuijsen, M.J., Colvile, R.N., 2001b. Determinants of fine particle (PM2.5) personal exposure levels in transport microenvironments, London, UK. Atmospheric Environment 35, 4557–4566.

Aethlabs, 2016. Aethlabs Microaetholometer Manual. URL: `https://aethlabs.com/sites/all/content/microaeth/microAethModelAE51OperatingManualRev05UpdatedMay2015.pdf`.

Allen, R., Adar, S., Avol, E., 2012. Modeling the residential infiltration of outdoor PM2. 5 in the Multi-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air). Environmental health perspectives 824, 824–830.

Armstrong, B.G., 1990. The effects of measurement errors on relative risk regressions. American journal of epidemiology 132, 1176–84.

Ashmore, M.R., Dimitroulopoulou, C., 2009. Personal exposure of children to air pollution. Atmospheric Environment 43, 128–141.

Ashworth, D.C., Fuller, G.W., Toledano, M.B., Font, A., Elliott, P., Hansell, A.L., de Hoogh, K., 2013. Comparative assessment of particulate air pollution exposure from municipal solid waste incinerator emissions. Journal of environmental and public health 2013, 560342.

Atkinson, R.W., Fuller, G.W., Anderson, H.R., Harrison, R.M., Armstrong, B., 2010. Urban ambient particle metrics and health: a time-series analysis. Epidemiology (Cambridge, Mass.) 21, 501–11.

Atkinson, R.W., Kang, S., Anderson, H.R., Mills, I.C., Walton, H.a., 2014. Epidemiological time series studies of PM2.5 and daily mortality and hospital admissions: a systematic review and meta-analysis. Thorax , 1–6.

Baumgartner, J., Schauer, J.J., Ezzati, M., Lu, L., Cheng, C., Patz, J., Bautista, L.E., 2011. Patterns and predictors of personal exposure to indoor air pollution from biomass combustion among women and children in rural China. Indoor air 21, 479–88.

Baxter, L.K., Barzyk, T.M., Vette, A.F., Croghan, C., Williams, R.W., 2008. Contributions of diesel truck emissions to indoor elemental carbon concentrations in homes in proximity to Ambassador Bridge. Atmospheric Environment 42, 9080–9086.

Baxter, L.K., Dionisio, K.L., Burke, J., Ebelt Sarnat, S., Sarnat, J.a., Hodas, N., Rich, D.Q., Turpin, B.J., Jones, R.R., Mannshardt, E., Kumar, N., Beevers, S.D., Özkaynak, H., 2013. Exposure prediction approaches used in air pollution epidemiology studies: key findings and future recommendations. Journal of exposure science & environmental epidemiology 23, 654–9.

BBC, 2007. Pollution risk for Olympic events.

Beelen, R., Raaschou-Nielsen, O., Stafoggia, M., Andersen, Z.J., Weinmayr, G., Hoffmann, B., Wolf, K., Samoli, E., Fischer, P., Nieuwenhuijsen, M., Vineis, P., Xun, W.W., Katsouyanni, K., Dimakopoulou, K., Oudin, A., Forsberg, B., Modig, L., Havulinna, A.S., Lanki, T., Turunen, A., Oftedal, B., Nystad, W., Nafstad, P., De Faire, U., Pedersen, N.L., Ostenson, C.G., Fratiglioni, L., Penell, J., Korek, M., Pershagen, G., Eriksen, K.T., Overvad, K., Ellermann, T., Eeftens, M., Peeters, P.H., Meliefste, K., Wang, M., Bueno-de Mesquita, B., Sugiri, D., Krämer, U., Heinrich, J., de Hoogh, K., Key, T., Peters, A., Hampel, R., Concin, H., Nagel, G., Ineichen, A., Schaffner, E., Probst-Hensch, N., Künzli, N., Schindler, C., Schikowski, T., Adam, M., Phuleria, H., Vilier, A., Clavel-Chapelon, F., Declercq, C., Grioni, S., Krogh, V., Tsai, M.Y., Ricceri, F., Sacerdote, C., Galassi, C., Migliore, E., Ranzi, A., Cesaroni, G., Badaloni, C., Forastiere, F., Tamayo, I., Amiano, P., Dorronsoro, M., Katsoulis, M., Trichopoulou, A., Brunekreef, B., Hoek, G., 2013. Effects of long-term exposure to air pollution on natural-cause mortality: an analysis of 22 European cohorts within the multicentre ESCAPE project. Lancet 6736, 1–11.

Beevers, S.D., Kitwiroon, N., Williams, M.L., Kelly, F.J., Ross Anderson, H., Carslaw, D.C., 2013. Air pollution dispersion models for human exposure predictions in London. Journal of exposure science & environmental epidemiology 23, 647–53.

Bell, M.L., Davis, D.L., Fletcher, T., 2003. A Retrospective Assessment of Mortality from the London Smog Episode of 1952: The Role of Influenza and Pollution. Environmental Health Perspectives 112, 6–8.

Beverland, I., Robertson, C., Yap, C., Heal, M., Cohen, G., Henderson, D., Hart, C., Agius, R., 2012a. Comparison of models for estimation of long-term exposure to air pollution in cohort studies. Atmospheric Environment 62, 530–539.

Beverland, I.J., Cohen, G.R., Heal, M.R., Carder, M., Yap, C., Robertson, C., Hart, C.L., Agius, R.M., 2012b. A comparison of short-term and long-term air pollution exposure associations with mortality in two cohorts in Scotland. Environmental health perspectives 120, 1280–5.

Boldo, E., Linares, C., Lumbreras, J., Borge, R., Narros, A., García-Pérez, J., Fernández-Navarro, P., Pérez-Gómez, B., Aragonés, N., Ramis, R., Pollán, M., Moreno, T., Karanasiou, A., López-Abente, G., 2011. Health impact assessment of a reduction in ambient PM(2.5) levels in Spain. Environment international 37, 342–8.

Bonita, R., Beaglehole, R., Kjellström, T., 2006. Basic epidemiology.

Branis, M., Safránek, J., Hytychová, A., 2009. Exposure of children to airborne particulate matter of different size fractions during indoor physical education at school. Building and Environment 44, 1246–1252.

Brauer, M., Amann, M., Burnett, R.T., Cohen, A., Dentener, F., Ezzati, M., Henderson, S.B., Krzyzanowski, M., Martin, R.V., Van Dingenen, R., van Donkelaar, A., Thurston, G.D., 2012. Exposure assessment for estimation of the global burden of disease attributable to outdoor air pollution. Environmental science & technology 46, 652–60.

Brauer, M., Brumm, J., Vedal, S., Petkau, a.J., 2002. Exposure misclassification and threshold concentrations in time series analyses of air pollution health effects. Risk analysis : an official publication of the Society for Risk Analysis 22, 1183–93.

Briggs, D., Collins, S., Elliott, P., Fischer, P.H., Kingham, S., Lebret, E., Pryl, K., Ree, V., Smallbone, K., Van Der Veen, A., 1997. Mapping urban air pollution using GIS: a regression-based approach. International Journal of Geographical Information Science 11, 699–718.

Brimblecombe, P., 1999. Air pollution and health history. Air pollution and health .

Britter, R., Hanna, S., 2003. Flow and dispersion in urban areas. Annual Review of Fluid Mechanics 35, 469–496.

Broday, D.M., Arpaci, A., Bartonova, A., Castell-Balaguer, N., Cole-Hunter, T., Dauge, F.R., Fishbain, B., Jones, R.L., Galea, K., Jovasevic-Stojanovic, M., Kocman, D., Martinez-Iñiguez, T., Nieuwenhuijsen, M., Robinson, J., Svecova, V., Thai, P., 2017. Wireless distributed environmental sensor networks for air pollution measurement-the promise and the current reality. Sensors (Switzerland) doi:10.3390/s17102263.

Broich, A.V., Gerharz, L.E., Klemm, O., 2011. Personal monitoring of exposure to particulate matter with a high temporal resolution. Environmental science and pollution research international 19, 2959–72.

Brook, R.D., Rajagopalan, S., Pope, C.A., Brook, J.R., Bhatnagar, A., Diez-Roux, A.V., Holguin, F., Hong, Y., Luepker, R.V., Mittleman, M.a., Peters, A., Siscovick, D., Smith, S.C., Whitsel, L., Kaufman, J.D., 2010. Particulate matter air pollution and cardiovascular disease: An update to the scientific statement from the American Heart Association. Circulation 121, 2331–78.

Brunekreef, B., 2007. Health effects of air pollution observed in cohort studies in Europe. Journal of exposure science & environmental epidemiology 17 Suppl 2, S61–S65.

Buonanno, G., Stabile, L., Morawska, L., 2014. Personal exposure to ultrafine particles: the influence of time-activity patterns. The Science of the total environment 468-469, 903–7.

Buteau, S., Hatzopoulou, M., Crouse, D.L., Smargiassi, A., Burnett, R.T., Logan, T., Cavellin, L.D., Goldberg, M.S., 2017. Comparison of spatiotemporal prediction models of daily exposure of individuals to ambient nitrogen dioxide and ozone in Montreal, Canada. Environmental Research 156, 201–230. doi:10.1016/j.envres.2017.03.017.

Cambridge Environmental Research Counsultants (CERC), 2014. ADMS-Urban.

Carslaw, D.C., ApSimon, H.M., Beevers, S.D., Brooks, D., Carruthers, D., Cooke, S., Kitwiroon, N., Oxley, T., Steadman, J., Stocker, J., 2013. Defra Phase 2 urban model evaluation. Technical Report October. King's College London. London. URL: http://uk-air.defra.gov.uk/assets/documents/reports/cat20/1312021020{_}131031urbanPhase2.pdf.

Carslaw, D.C., Beevers, S.D., Tate, J.E., Westmoreland, E.J., Williams, M.L., 2011. Recent evidence concerning higher NOx emissions from passenger cars and light duty vehicles. Atmospheric Environment 45, 7053–7063.

Chaix, B., Méline, J., Duncan, S., Merrien, C., Karusisi, N., Perchoux, C., Lewin, A., Labadi, K., Kestens, Y., 2013. GPS tracking in neighborhood and health studies: a step forward for environmental exposure assessment, a step backward for causal inference? Health & place 21, 46–51.

Challoner, A., Gill, L., 2014. Indoor/outdoor air pollution relationships in ten commercial buildings: PM2.5 and NO2. Building and Environment 80, 159–173.

Chang, W., Cheng, J., Allaire, J., Xie, Y., McPherson, J., 2018. shiny: Web Application Framework for R. URL: https://CRAN.R-project.org/package=shiny. r package version 1.1.0.

Chen, C., Zhao, B., 2011. Review of relationship between indoor and outdoor particles: I/O ratio, infiltration factor and penetration factor. Atmospheric Environment 45, 275–288.

Cheng, Y.H., Lin, M.H., 2013. Real-time performance of the microaeth?? AE51 and the effects of aerosol loading on its measurement results at a traffic site. Aerosol and Air Quality Research 13, 1853–1863. URL: http://www.aaqr.org/files/article/935/23{_}AAQR-12-12-OA-0371{_}1853-1863.pdf, doi:10.4209/aaqr.2012.12.0371.

CMAS Centre, 2014. CMAQ.

Colbeck, I., Nasir, Z.A., 2010. Human Exposure to Pollutants via Dermal Absorption and Inhalation. volume 17 of *Environmental Pollution*. Springer Netherlands, Dordrecht.

Colls, J., 1997. Air Pollution: An Introduction. E & FN Spon.

Committee on the Medical Effects of Air Pollutants, 2009. Long-Term Exposure to Air Pollution: Effect on Mortality. Technical Report. Committee on the Medical Effects of Air Pollutants. London.

Committee on the Medical Effects of Air Pollutants, 2010. The Mortality Effects of Long-Term Exposure to Particulate Air Pollution in the United Kingdom. Technical Report. Committee on the Medical Effects of Air Pollutants. London.

Committee on the Medical Effects of Air Pollutants, 2018. Committee on the Medical Effects of Air Pollutants Statement on quantifying mortality associated with long-term average concentrations of fine particulate matter (PM 2.5). Technical Report. Committee on the Medical Effects of Air Pollutants. URL: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment{_}data/file/734813/COMEAP{_}PM{_}2.5{_}statement.pdf.

Cyrys, J., Pitz, M., Heinrich, J., Wichmann, H.E., Peters, A., 2008. Spatial and temporal variation of particle number concentration in Augsburg, Germany. Science of the Total ... 401, 168–175.

DEFRA, 2007. The Air Quality Strategy for England, Scotland, Wales and Northern Ireland. Technical Report. DEFRA.

DEFRA, 2009. LAQM - Technical Guidance 09. Technical Report February. DEFRA. London.

DEFRA, 2011a. Causes of air pollution.

DEFRA, 2011b. Monitoring Networks - A Brief history.

Dhondt, S., Beckx, C., Degraeuwe, B., Lefebvre, W., Kochan, B., Bellemans, T., Int Panis, L., Macharis, C., Putman, K., 2012. Health impact assessment of air pollution using a dynamic exposure profile: Implications for exposure and health impact estimates. Environmental Impact Assessment Review 36, 42–51.

Dockery, D., Pope, C., Xu, X., 1993. An association between air pollution and mortality in six US cities. New England journal .

van Donkelaar, A., Martin, R.V., Brauer, M., Boys, B.L., 2015. Use of Satellite Observations for Long-Term Exposure Assessment of Global Concentrations of Fine Particulate Matter. Environmental Health Perspectives 123, 135–143. URL: https://ehp.niehs.nih.gov/doi/10.1289/ehp.1408646, doi:10.1289/ehp.1408646.

Dons, E., Int Panis, L., Van Poppel, M., Theunis, J., Willems, H., Torfs, R., Wets, G., 2011. Impact of timeactivity patterns on personal exposure to black carbon. Atmospheric Environment 45, 3594–3602.

Dons, E., Temmerman, P., Van Poppel, M., Bellemans, T., Wets, G., Int Panis, L., 2013. Street characteristics and traffic factors determining road users' exposure to black carbon. The Science of the total environment 447C, 72–79.

Environmental Protection Agency, 2008. Dispersion Modeling.

Environmental Protection Agency, 2009. Integrated Science Assessment for Particulate Matter. Technical Report. Environmental Protection Agency. Washington.

Environmental Protection Agency, 2012. Ground Level Ozone ¿ Basic Information.

Faustini, A., Rapp, R., Forastiere, F., 2014. Nitrogen dioxide and mortality: review and meta-analysis of long-term studies. The European respiratory journal , 1–10.

Ferro, A.R., Kopperud, R.J., Hildemann, L.M., 2004. Elevated personal exposure to particulate matter from human activities in a residence. Journal of Exposure Analysis and Environmental Epidemiology 14, 34–40. doi:10.1038/sj.jea.7500356.

Funasaka, K., Miyazaki, T., Tsuruho, K., Tamura, K., Mizuno, T., Kuroda, K., 2000. Relationship between indoor and outdoor carbonaceous particulates in roadside households. Environmental Pollution 110, 127–134.

Garza, G., 1996. Uncontrolled air pollution in Mexico City. Cities 13.

Gauderman, W.J., Vora, H., McConnell, R., Berhane, K., Gilliland, F., Thomas, D., Lurmann, F., Avol, E., Kunzli, N., Jerrett, M., Peters, J., 2007. Effect of exposure to traffic on lung development from 10 to 18 years of age: a cohort study. Lancet 369, 571–7.

Gens, A., Hurley, J.F., Tuomisto, J.T., Friedrich, R., 2014. Health impacts due to personal exposure to fine particles caused by insulation of residential buildings in Europe. Atmospheric Environment 84, 213–221.

Gerharz, L.E., Klemm, O., Broich, A.V., Pebesma, E., 2013. Spatio-temporal modelling of individual exposure to air pollution and its uncertainty. Atmospheric Environment 64, 56–65.

Ghio, A.J., Kim, C., Devlin, R.B., 2000. Concentrated ambient air particles induce mild pulmonary inflammation in healthy human volunteers. American journal of respiratory and critical care medicine 162, 981–8.

Global Health Observatory, W.H.O., 2012. Urban population growth.

Godfrey, A.J.R., Govindaraju, K., 2018. Dodge: Functions for acceptance sampling iIdeas originated by H.F. Dodge. URL: `https://cran.r-project.org/package=Dodge`. r package version 0.9-2.

Goldstein, I.F., Landovitz, L., 1977. Analysis of air pollution patterns in New York CityI. Can one station represent the large metropolitan area? Atmospheric Environment (1967) 11, 47–52.

Greater London Authority (GLA), 2002. 50 Years on: The Struggle for Air Quality in London Since the Great Smog of December 1952. Technical Report December. Greater London Authority (GLA). London.

Greater London Authority (GLA), 2010. The Mayors Air Quality Strategy : Clearing the air (Executive Summary). Technical Report December. Greater London Authority (GLA). London.

Gulliver, J., Briggs, D., 2004. Personal exposure to particulate air pollution in transport microenvironments. Atmospheric Environment 38, 1–8.

Gulliver, J., Briggs, D.J., 2007. Journey-time exposure to particulate air pollution. Atmospheric Environment 41, 7195–7207.

Han, Y., Qi, M., Chen, Y., Shen, H., Liu, J., Huang, Y., Chen, H., Liu, W., Wang, X., Liu, J., Xing, B., Tao, S., 2015. Influences of ambient air PM2.5 concentration and meteorological condition on the indoor PM2.5 concentrations in a residential apartment

in Beijing using a new approach. Environmental Pollution 205, 307–314. doi:10.1016/j.envpol.2015.04.026.

Hansen, A., Rosen, H., Novakov, T., 1984. The aethalometer An instrument for the real-time measurement of optical absorption by aerosol particles. Science of The Total Environment 36, 191–196. URL: https://www.sciencedirect.com/science/article/pii/0048969784902651http://linkinghub.elsevier.com/retrieve/pii/0048969784902651, doi:10.1016/0048-9697(84)90265-1.

Health Effects Institute, 2010. Traffic-related air pollution: a critical review of the literature on emissions, exposure, and health effects. Health Effects Institute, Boston .

Hoek, G., 2017. Methods for Assessing Long-Term Exposures to Outdoor Air Pollutants. Current Environmental Health Reports 4, 450–462. URL: http://link.springer.com/10.1007/s40572-017-0169-5, doi:10.1007/s40572-017-0169-5.

Hoek, G., Beelen, R., de Hoogh, K., Vienneau, D., Gulliver, J., Fischer, P., Briggs, D., 2008. A review of land-use regression models to assess spatial variation of outdoor air pollution. Atmospheric Environment 42, 7561–7578.

Hoek, G., Krishnan, R.M., Beelen, R., Peters, A., Ostro, B., Brunekreef, B., Kaufman, J.D., 2013. Long-term air pollution exposure and cardio- respiratory mortality: a review. Environmental health : a global access science source 12, 43.

Huang, L., Hopke, P.K., Zhao, W., Li, M., 2015. Determinants on ambient PM2.5 infiltration in non-heating season for urban residences in Beijing: Building characteristics, interior surface coverings and human behavior. Atmospheric Pollution Research 6, 1046–1054. doi:10.1016/j.apr.2015.05.009.

Hurley, J.F., Cherrie, J.W., Donaldson, K., Seaton, A., Tran, C.L., 2003. Assessment of health effects of long-term occupational exposure to tunnel dust in the London Underground. Technical Report December. Institute of occupational Medicine.

Hussein, T., Wierzbicka, A., Löndahl, J., Lazaridis, M., Hänninen, O., 2014. Indoor aerosol modeling for assessment of exposure and respiratory tract deposited dose. Atmospheric Environment .

Jalava, P.I., Aakko-Saksa, P., Murtonen, T., Happo, M.S., Markkanen, A., Yli-Pirilä, P., Hakulinen, P., Hillamo, R., Mäki-Paakkanen, J., Salonen, R.O., Jokiniemi, J., Hirvonen, M.R., 2012. Toxicological properties of emission particles from heavy duty engines powered by conventional and bio-based diesel fuels and compressed natural gas. Particle and fibre toxicology 9, 37.

Janssen, N.A.H., van Vliet, P.H.N., Aarts, F., Harssema, H., Brunekreef, B., 2001. Assessment of exposure to traffic related air pollution of children attending schools near motorways. Atmospheric Environment 35, 3875–3884.

Jiang, R.T., Acevedo-Bolton, V., Cheng, K.C., Klepeis, N.E., Ott, W.R., Hildemann, L.M., 2011. Determination of response of real-time SidePak AM510 monitor to secondhand smoke, other common indoor aerosols, and outdoor aerosol. Journal of environmental monitoring : JEM 13, 1695–702. URL: http://www.ncbi.nlm.nih.gov/pubmed/21589975, doi:10.1039/c0em00732c.

Karanasiou, A., Viana, M., Querol, X., Moreno, T., de Leeuw, F., 2014. Assessment of personal exposure to particulate air pollution during commuting in European cities–recommendations and policy implications. The Science of the total environment 490, 785–97.

Kaur, S., Nieuwenhuijsen, M., Colvile, R., 2005. Personal exposure of street canyon intersection users to PM2.5, ultrafine particle counts and carbon monoxide in Central London, UK. Atmospheric Environment 39, 3629–3641.

Kearney, J., Wallace, L., MacNeill, M., Héroux, M.E., Kindzierski, W., Wheeler, A., 2014. Residential infiltration of fine and ultrafine particles in Edmonton. Atmospheric Environment 94, 793–805.

King's College London, 2013. London Air Quality Network.

Kloog, I., Coull, B.a., Zanobetti, A., Koutrakis, P., Schwartz, J.D., 2012. Acute and chronic effects of particles on hospital admissions in New-England. PloS one 7, e34664.

Kloog, I., Ridgway, B., Koutrakis, P., Coull, B.a., Schwartz, J.D., 2013. Long- and short-term exposure to PM2.5 and mortality: using novel exposure models. Epidemiology (Cambridge, Mass.) 24, 555–61.

Knibbs, L.D., Cole-Hunter, T., Morawska, L., 2011. A review of commuter exposure to ultrafine particles and its health effects. Atmospheric Environment 45, 2611–2622.

Kousa, A., Kukkonen, J., Karppinen, A., Aarnio, P., Koskentalo, T., 2002. A model for evaluating the population exposure to ambient air pollution in an urban area. Atmospheric Environment 36, 2109–2119.

Lai, H.K., Kendall, M., Ferrier, H., Lindup, I., Alm, S., Hänninen, O., Jantunen, M., Mathys, P., Colvile, R., Ashmore, M.R., Cullinan, P., Nieuwenhuijsen, M.J., 2004. Personal exposures and microenvironment concentrations of PM2.5, VOC, NO2 and CO in Oxford, UK. Atmospheric Environment 38, 6399–6410.

Lao, J., Teixid, O., 2011. Air quality model for Barcelona. WIT Transactions on Ecology and the Environment 147, 25–36. doi:10.2495/AIR110031.

Larsen, J., 2013. Bike-Sharing Programs Hit the Streets in Over 500 Cities Worldwide. Earth Policy Institute 25.

Lee, J., Lim, S., Lee, K., Guo, X., Kamath, R., Yamato, H., Abas, A.L., Nandasena, S., Nafees, A.A., Sathiakumar, N., 2010. Secondhand smoke exposures in indoor public places in seven Asian countries. International Journal of Hygiene and Environmental Health 213, 348–351.

Li, Z., Sjödin, A., Romanoff, L.C., Horton, K., Fitzgerald, C.L., Eppler, A., Aguilar-Villalobos, M., Naeher, L.P., 2011. Evaluation of exposure reduction to indoor air pollution in stove intervention projects in Peru by urinary biomonitoring of polycyclic aromatic hydrocarbon metabolites. Environment International 37, 1157–1163.

for London, T., 2018. Travel in London Report 11. Technical Report. Transport for London. URL: http://content.tfl.gov.uk/travel-in-london-report-11.pdf.

Loomis, D., Grosse, Y., Lauby-Secretan, B., El Ghissassi, F., Bouvard, V., Benbrahim-Tallaa, L., Guha, N., Baan, R., Mattock, H., Straif, K., Ghissassi, F.E., 2013. The carcinogenicity of outdoor air pollution. The Lancet Oncology 14, 1262–1263.

Lovelace, R., Ellison, R., 2018. stplanr: Sustainable Transport Planning. URL: https://CRAN.R-project.org/package=stplanr. r package version 0.2.5.

Loxham, M., Cooper, M.J., Gerlofs-Nijland, M.E., Cassee, F.R., Davies, D.E., Palmer, M.R., Teagle, D.a.H., 2013. Physicochemical characterization of airborne particulate matter at a mainline underground railway station. Environmental science & technology 47, 3614–22.

MacNeill, M., Wallace, L., Kearney, J., Allen, R., Van Ryswyk, K., Judek, S., Xu, X., Wheeler, A., 2012. Factors influencing variability in the infiltration of PM2.5 mass and its components. Atmospheric Environment 61, 518–532.

Mannino, D., 2000. What constitutes an adverse health effect of air pollution? American Journal of Respiratory and Critical Care . . . .

Maroko, A.R., 2012. Using air dispersion modeling and proximity analysis to assess chronic exposure to fine particulate matter and environmental justice in New York City. Applied Geography 34, 533–547.

Mayer, H., 1999. Air pollution in cities. Atmospheric Environment 33, 4029–4037.

Mazzi, E.a., Dowlatabadi, H., 2007. Air quality impacts of climate mitigation: UK policy and passenger vehicle choice. Environmental science & technology 41, 387–92.

Meliker, J.R., Sloan, C.D., 2011. Spatio-temporal epidemiology: principles and opportunities. Spatial and spatio-temporal epidemiology 2, 1–9.

Miller, B., 2010. Report on estimation of mortality impacts of particulate air pollution in London. Institute of Occupational Medicine (IOM) .

Minguillón, M., Schembari, A., Triguero-Mas, M., de Nazelle, A., Dadvand, P., Figueras, F., Salvado, J., Grimalt, J., Nieuwenhuijsen, M., Querol, X., 2012. Source apportionment of indoor, outdoor and personal PM2.5 exposure of pregnant women in Barcelona, Spain. Atmospheric Environment 59, 426–436.

Moltchanov, S., Levy, I., Etzion, Y., Lerner, U., Broday, D.M., Fishbain, B., 2015. On the feasibility of measuring urban air pollution by wireless distributed sensor networks. Science of the Total Environment doi:10.1016/j.scitotenv.2014.09.059.

Mölter, A., Lindley, S., de Vocht, F., Agius, R., Kerry, G., Johnson, K., Ashmore, M., Terry, A., Dimitroulopoulou, S., Simpson, A., 2012. Performance of a microenviromental model for estimating personal NO2 exposure in children. Atmospheric Environment 51, 225–233.

Monks, P.S., Granier, C., Fuzzi, S., Stohl, A., Williams, M.L., Akimoto, H., Amann, M., Baklanov, A., Baltensperger, U., Bey, I., Blake, N., Blake, R.S., Carslaw, K., Cooper, O.R., Dentener, F., Fowler, D., Fragkou, E., Frost, G.J., Generoso, S., Ginoux, P., Grewe, V., Guenther, A., Hansson, H.C., Henne, S., Hjorth, J., Hofzumahaus, A., Huntrieser, H., Isaksen, I.S.A., Jenkin, M.E., Kaiser, J., Kanakidou, M., Klimont, Z., Kulmala, M., Laj, P., Lawrence, M.G., Lee, J.D., Liousse, C., Maione, M., McFiggans, G., Metzger, A., Mieville, A., Moussiopoulos, N., Orlando, J.J., O'Dowd, C.D., Palmer, P.I., Parrish, D.D., Petzold, A., Platt, U., Pöschl, U., Prévôt, A.S.H., Reeves, C.E., Reimann, S., Rudich, Y., Sellegri, K., Steinbrecher, R., Simpson, D., ten Brink, H., Theloke, J., van der Werf, G.R., Vautard, R., Vestreng, V., Vlachokostas, C., von Glasow, R., 2009. Atmospheric composition change - global and regional air quality. Atmospheric Environment 43, 5268–5350.

Moore, E., Chatzidiakou, L., Jones, R.L., Smeeth, L., Beevers, S., Kelly, F.J., K Quint, J., Barratt, B., 2016. Linking e-health records, patient-reported symptoms and environmental exposure data to characterise and model COPD exacerbations: protocol for the COPE study. BMJ open 6, e011330. URL: http://www.ncbi.nlm.nih.gov/pubmed/27412104http:

//www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4947745, doi:10.1136/bmjopen-2016-011330.

Nasir, Z.A., Colbeck, I., 2009. Particulate air pollution in transport micro-environments. Journal of environmental monitoring : JEM 11, 1140–6.

de Nazelle, A., Basagana, X., Figueras, F., Sunyer, J., Nieuwenhuijsen, M., 2008. Air Pollution Exposure and Reproductive Outcomes Study in Barcelona. Epidemiology 19.

de Nazelle, A., Seto, E., Donaire-Gonzalez, D., Mendez, M., Matamala, J., Nieuwenhuijsen, M.J., Jerrett, M., 2013. Improving estimates of air pollution exposure through ubiquitous sensing technologies. Environmental pollution (Barking, Essex : 1987) 176, 92–9.

Nieuwenhuijsen, M.J., Gómez-Perales, J.E., Colvile, R.N., 2007. Levels of particulate air pollution, its elemental composition, determinants and health effects in metro systems. Atmospheric Environment 41, 7995–8006. URL: http://www.sciencedirect.com/science/article/pii/S135223100700698X, doi:10.1016/j.atmosenv.2007.08.002.

Nwokoro, C., Ewin, C., Harrison, C., Ibrahim, M., Dundas, I., Dickson, I., Mushtaq, N., Grigg, J., 2012. Cycling to work in London and inhaled dose of black carbon. The European respiratory journal 40, 1091–7.

Nyhan, M., Grauwin, S., Britter, R., Misstear, B., McNabola, A., Laden, F., Barrett, S.R.H., Ratti, C., 2016. Exposure TrackThe Impact of Mobile-Device-Based Mobility Patterns on Quantifying Population Exposure to Air Pollution. Environmental Science & Technology 50, 9671–9681. URL: http://pubs.acs.org/doi/10.1021/acs.est.6b02385, doi:10.1021/acs.est.6b02385.

Office for National Statistics, 2010. Social Trends, No. 40, 2010 Edition. Technical Report. Office for National Statistics.

Office for National Statistics, 2014. 2011 Census Analysis, Cycling to Work.

Ordnance Survey, 2015a. Code-Point Open.

Ordnance Survey, 2015b. Code-Point with polygons.

Özkaynak, H., Baxter, L.K., Dionisio, K.L., Burke, J., 2013. Air pollution exposure prediction approaches used in air pollution epidemiology studies. Journal of exposure science & environmental epidemiology 23, 566–72.

Peacock, J.L., Anderson, H.R., Bremner, S.a., Marston, L., Seemungal, T.a., Strachan, D.P., Wedzicha, J.a., 2011. Outdoor air pollution and respiratory health in patients with COPD. Thorax 66, 591–6.

Pedersen, M., Giorgis-allemand, L., Bernard, C., Aguilera, I., Andersen, A.m.N., Ballester, F., Beelen, R.M.J., 2013. Articles Ambient air pollution and low birthweight : a European cohort study (ESCAPE). Lancet 2600.

PennState Eberly College of Science, 2017. 8.2.1.3 - Computing Necessary Sample Size — STAT 200. URL: https://onlinecourses.science.psu.edu/stat200/node/256.

Pfeifer, G.D., Harrison, R.M., Lynam, D.R., 1999. Personal exposures to airborne metals in London taxi drivers and office workers in 1995 and 1996. The Science of the total environment 235, 253–60. URL: http://www.ncbi.nlm.nih.gov/pubmed/10535124.

Pope III, C.A., Iii, C.A.P., Burnett, R.T., Thun, M.J., Calle, E.E., Krewski, D., Thurston, G.D., 2012. Lung Cancer, Cardiopulmonary Mortality, and Long-term Exposure to Fine Particulate Air Pollution. Jama The Journal Of The American Medical Association 287, 1132–1141.

Ragettli, M.S., Corradi, E., Braun-Fahrländer, C., Schindler, C., de Nazelle, A., Jerrett, M., Ducret-Stich, R.E., Künzli, N., Phuleria, H.C., 2013. Commuter exposure to ultrafine particles in different urban locations, transportation modes and routes. Atmospheric Environment 77, 376–384.

Reis, S., Liška, T., Vieno, M., Carnell, E.J., Beck, R., Clemens, T., Dragosits, U., Tomlinson, S.J., Leaver, D., Heal, M.R., 2018. The influence of residential and workday population mobility on exposure to air pollution in the UK. Environment International 121, 803–813. URL: https://www.sciencedirect.com/science/article/pii/S016041201830864X, doi:10.1016/J.ENVINT.2018.10.005.

Restrepo, C., Zimmerman, R., Thurston, G., Clemente, J., Gorczynski, J., Zhong, M., Blaustein, M., Chi Chen, L., 2004. A comparison of ground-level air quality data with New York State Department of Environmental Conservation monitoring stations data in South Bronx, New York. Atmospheric Environment 38, 5295–5304.

Rojas-Rueda, D., de Nazelle, A., Teixidó, O., Nieuwenhuijsen, M.J., 2013. Health impact assessment of increasing public transport and cycling use in Barcelona: a morbidity and burden of disease approach. Preventive medicine 57, 573–9.

Rose, N., Cowie, C., Gillett, R., Marks, G.B., 2009. Weighted road density: A simple way of assigning traffic-related air pollution exposure. Atmospheric Environment 43, 5009–5014.

Russell, W., 1926. THE RELATIVE INFLUENCE OF FOG AND LOW TEMPERATURE ON THE MORTALITY FROM RESPIRATORY DISEASE. The Lancet , 335–339.

Salvi, S., Blomberg, A., Rudell, B., Kelly, F., Sandström, T., Holgate, S.T., Frew, A., 1999. Acute inflammatory responses in the airways and peripheral blood after short-term exposure to diesel exhaust in healthy human volunteers. American journal of respiratory and critical care medicine 159, 702–9.

Salvi, S.S., Nordenhall, C., Blomberg, A., Rudell, B., Pourazar, J., Kelly, F.J., Wilson, S., Sandström, T., Holgate, S.T., Frew, a.J., 2000. Acute exposure to diesel exhaust increases IL-8 and GRO-alpha production in healthy human airways. American journal of respiratory and critical care medicine 161, 550–7.

Samoli, E., Analitis, A., Touloumi, G., Schwartz, J., Anderson, H.R., Sunyer, J., Bisanti, L., Zmirou, D., Vonk, J.M., Pekkanen, J., Goodman, P., Paldy, A., Schindler, C., Katsouyanni, K., 2004. Estimating the ExposureResponse Relationships between Particulate Matter and Mortality within the APHEA Multicity Project. Environmental Health Perspectives 113, 88–95.

Schweizer, C., Edwards, R.D., Bayer-Oglesby, L., Gauderman, W.J., Ilacqua, V., Jantunen, M.J., Lai, H.K., Nieuwenhuijsen, M., Künzli, N., 2007. Indoor time-microenvironment-activity patterns in seven regions of Europe. Journal of exposure science & environmental epidemiology 17, 170–81.

Scoggins, A., Kjellstrom, T., Fisher, G., Connor, J., Gimson, N., 2004. Spatial analysis of annual air pollution exposure and mortality. The Science of the total environment 321, 71–85.

Seaton, a., Cherrie, J., Dennekamp, M., Donaldson, K., Hurley, J.F., Tran, C.L., 2005. The London Underground: dust and hazards to health. Occupational and environmental medicine 62, 355–62.

Seneca, L.A., Campbell, R., 1969. Epistulae Morales Ad Lucilium:. Classics Series, Penguin Books Limited.

Shi, Y., Lau, K.K.L., Ng, E., 2016. Developing Street-Level PM2.5 and PM10 Land Use Regression Models in High-Density Hong Kong with Urban Morphological Factors. Environmental Science and Technology 50, 8178–8187. doi:10.1021/acs.est.6b01807.

Silva, R.a., West, J.J., Zhang, Y., Anenberg, S.C., Lamarque, J.F., Shindell, D.T., Collins, W.J., Dalsoren, S., Faluvegi, G., Folberth, G., Horowitz, L.W., Nagashima, T., Naik, V., Rumbold, S., Skeie, R., Sudo, K., Takemura, T., Bergmann, D., Cameron-Smith, P.,

Cionni, I., Doherty, R.M., Eyring, V., Josse, B., MacKenzie, I.a., Plummer, D., Righi, M., Stevenson, D.S., Strode, S., Szopa, S., Zeng, G., 2013. Global premature mortality due to anthropogenic outdoor air pollution and the contribution of past climate change. Environmental Research Letters 8, 034005.

Slezak, M., 2013. China plans to create artificial rain to combat smog. New Scientist 220, 7.

Smethurst, H., Witham, C., Robins, A., Murray, V., 2012. An exceptional case of long range odorant transport. Journal of Wind Engineering and Industrial Aerodynamics 103, 60–72.

Smith, J.D., Mitsakou, C., Kitwiroon, N., Barratt, B.M., Walton, H.A., Taylor, J.G., Anderson, H.R., Kelly, F.J., Beevers, S.D., 2016. The London Hybrid Exposure Model (LHEM): Improving human exposure estimates to NO ₂ and PM _{2.5} in an urban setting. Environmental Science & Technology , acs.est.6b01817URL: http://pubs.acs.org/doi/abs/10.1021/acs.est.6b01817, doi:10.1021/acs.est.6b01817.

Song, W.W., Ashmore, M.R., Terry, A.C., 2009. The influence of passenger activities on exposure to particles inside buses. Atmospheric Environment 43, 6271–6278.

Steinle, S., Reis, S., Sabel, C.E., 2013. Quantifying human exposure to air pollution-Moving from static monitoring to spatio-temporally resolved personal exposure assessment. The Science of the total environment 443, 184–93.

Stieb, D.M., Judek, S., Burnett, R.T., 2002. Meta-Analysis of Time-Series Studies of Air Pollution and Mortality: Effects of Gases and Particles and the Influence of Cause of Death, Age, and Season. Journal of the Air & Waste Management Association 52, 470–484.

Stohl, a., 2003. A backward modeling study of intercontinental pollution transport using aircraft measurements. Journal of Geophysical Research 108, 4370.

Summers, P., Whelpdale, D., 1976. Acid precipitation in Canada. Water, Air, and Soil Pollution 6.

SUN, Y., ZHUANG, G., WANG, Y., HAN, L., GUO, J., DAN, M., ZHANG, W., WANG, Z., HAO, Z., 2004. The air-borne particulate pollution in Beijing?concentration, composition, distribution and sources. Atmospheric Environment 38, 5991–6004.

Taylor, J., Shrubsole, C., Davies, M., Biddulph, P., Das, P., Hamilton, I., Vardoulakis, S., Mavrogianni, a., Jones, B., Oikonomou, E., 2014. The modifying effect of the building envelope on population exposure to PM2.5 from outdoor sources. Indoor air , 1–13.

The Guardian, 2008. What's that smell?

The United Nations Statistics Division, 2013. UNData - Country Profiles.

The World Bank, 2013. The World Band.

ThermoFisher Scientific, 2016. Partisol 2025i Sequential Air Sampler. URL: https://www.thermofisher.com/order/catalog/product/2025I.

Tonne, C., Beevers, S., Kelly, F.J., Jarup, L., Wilkinson, P., Armstrong, B., 2010. An approach for estimating the health effects of changes over time in air pollution: an illustration using cardio-respiratory hospital admissions in London. Occupational and environmental medicine 67, 422–7.

Tonne, C., Milà, C., Fecht, D., Alvarez, M., Gulliver, J., Smith, J., Beevers, S., Ross Anderson, H., Kelly, F., 2018. Socioeconomic and ethnic inequalities in exposure to air and noise pollution in London. Environment International 115, 170–179. URL: https://www.sciencedirect.com/science/article/pii/S0160412017321256, doi:10.1016/J.ENVINT.2018.03.023.

Torrey, C.M., Moon, K.A., Williams, D.A.L., Green, T., Cohen, J.E., Navas-Acien, A., Breysse, P.N., 2015. Waterpipe cafes in Baltimore, Maryland: Carbon monoxide, particulate matter, and nicotine exposure. Journal of exposure science & environmental epidemiology 25, 405–10. URL: http://www.ncbi.nlm.nih.gov/pubmed/24736103http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4333110, doi:10.1038/jes.2014.19.

Transport for London, 2014a. Barclays Cycle Hire.

Transport for London, 2014b. TfL - Facts and Figures.

Transport for London, 2016. Transport for London: Rolling Stock. URL: https://tfl.gov.uk/corporate/about-tfl/what-we-do/london-underground/rolling-stock.

TSI, 2015. SidePak Personal Aerosol Monitor AM510. URL: http://www.tsi.com/SIDEPAK-Personal-Aerosol-Monitor-AM510/.

Tuo, Y., Li, X., Wang, J., 2013. Negative Effects of Beijing's Air Pollution Caused by Urbanization on Residents Health, in: Proceedings of the 2nd International Conference on Science and Social Research, Atlantis Press, Paris, France. pp. 732–735.

United Nations, 1998. Kyoto Protocol to the United Nations Framework Convention on Climate Change.

United Nations Economic Commission for Europe, 1983. 1979 Convention on Long-Range Transboundary Air Pollution.

United States Environmental Protection Agency, 2008. Care for Your Air: A Guide to Indoor Air Quality. Technical Report September. United States Environmental Protection Agency.

United States Environmental Protection Agency, 2014. Community Multi-scale Air Quality Model.

U.S. Department of Health & Human Services, 2014. What is Epidemiology?

Viana, M., Rivas, I., Reche, C., Fonseca, A.S., Pérez, N., Querol, X., Alastuey, A., Álvarez-Pedrerol, M., Sunyer, J., 2015. Field comparison of portable and stationary instruments for outdoor urban air exposure assessments. Atmospheric Environment 123, 220–228. URL: http://linkinghub.elsevier.com/retrieve/pii/S1352231015304921, doi:10.1016/j.atmosenv.2015.10.076.

Walton, H., Dajnak, D., Beevers, S., Williams, M., Watkiss, P., Hunt, A., 2015. Understanding the health impacts of air pollution in london. London: Kings College London, Transport for London and the Greater London Authority .

Wang, C., Tu, Y., Yu, Z., Lu, R., 2015. PM2.5 and Cardiovascular Diseases in the Elderly: An Overview. International journal of environmental research and public health 12, 8187–97. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4515716{&}tool=pmcentrez{&}rendertype=abstract, doi:10.3390/ijerph120708187.

Wang, T., Xie, S., 2009. Assessment of traffic-related air pollution in the urban streets before and during the 2008 Beijing Olympic Games traffic control period. Atmospheric Environment 43, 5682–5690.

Wikipedia, 2014. List of bicycle sharing systems.

Williams, M., 2013. Failure to test for the real world has left a polluting diesel legacy.

Willocks, L.J., Bhaskar, A., Ramsay, C.N., Lee, D., Brewster, D.H., Fischbacher, C.M., Chalmers, J., Morris, G., Scott, E.M., 2012. Cardiovascular disease and air pollution in Scotland: no association or insufficient data and study design? BMC Public Health 12, 227. URL: http://bmcpublichealth.biomedcentral.com/articles/10.1186/1471-2458-12-227, doi:10.1186/1471-2458-12-227.

Woodcock, J., Tainio, M., Cheshire, J., O'Brien, O., Goodman, A., 2014. Health effects of the London bicycle sharing system: health impact modelling study. Bmj 348, g425–g425.

World Health Organisation, 2006. Air Quality Guidelines: Global Update 2005: Particulate Matter, Ozone, Nitrogen Dioxide and Sulfur Dioxide.

World Health Organization, 2005. Principles of Characterizing and Applying Human Exposure Models. Technical Report 3. World Health Organization.

World Health Organization, 2010. WHO guidelines for indoor air quality: selected pollutants. Geneva, Switzerland: World Health Organization .

World Health Organization, 2011. WHO — Air quality and health.

World Health Organization, 2012. Database: outdoor air pollution in cities.

World Health Organization, 2013a. Health Effects of Particulate Matter. Technical Report. World Health Organization.

World Health Organization, 2013b. Review of evidence on health aspects of air pollution REVIHAAP Project. Technical Report. World Health Organization.

Yim, S.H.L., Barrett, S.R.H., 2012. Public health impacts of combustion emissions in the United Kingdom. Environmental science & technology 46, 4291–6.

Yu, C.H., Fan, Z., Lioy, P.J., Baptista, A., Greenberg, M., Laumbach, R.J., 2016. A novel mobile monitoring approach to characterize spatial and temporal variation in traffic-related air pollutants in an urban community. Atmospheric Environment 141, 161–173. doi:10.1016/j.atmosenv.2016.06.044.

Zeger, S.L., Thomas, D., Dominici, F., Samet, J.M., Schwartz, J., Dockery, D., Cohen, a., 2000. Exposure measurement error in time-series studies of air pollution: concepts and consequences. Environmental health perspectives 108, 419–26.

Zuurbier, M., Hoek, G., Oldenwening, M., Lenters, V., Meliefste, K., van den Hazel, P., Brunekreef, B., 2010. Commuters' exposure to particulate matter air pollution is affected by mode of transport, fuel type, and route. Environmental health perspectives 118, 783–9.

Glossary

ADMS-Urban Atmospheric Dispersion Modelling System. 123

AOD Aerosol Optical Depth. 45

API Application Programming Interface. 11, 15, 100, 117–119, 215

AURN Automatic Urban and Rural Network. 47

BBC British Broadcasting Corporation. 29

CMAQ Community Multi-scale Air Quality model. 1, 13, 16, 87, 122, 123, 181–184, 186–189, 191, 195–198, 203–206, 210, 214, 216

CO Carbon monoxide. 20, 26, 27, 187

CO Nitrogen oxide. 10, 30, 31, 47, 123

CO_2 Carbon dioxide. 31

COMEAP Committee on Medical Effects of Air Pollutants. 36, 37, 41, 42, 211

COPD Chronic Obstructive Pulmonary Disorder. 175, 210

CSV comma-seperated-values. 95, 158, 159, 196

DALY Disability-Adjusted Life Year. 35, 39, 72

DEFRA Department for Environment, Food and Rural Affairs. 19, 20, 125

EC European Commission. 34

ERG Environmental Research Group. 123

EU European Union. 31, 47, 48, 61, 188

GIS Geographic Information Systems. 80, 87, 205, 207

GLA Greater London Authority. 210, 217

GPS Global Positioning System. 13, 14, 56, 58, 60, 82–84, 197, 205

HIA Health Impact Assessment. 72

IARC International Agency for Research on Cancer. 35

KCL King's College London. 123, 156, 176, 210, 211

LAEI London Atmospheric Emissions Inventory. 210

LAQN London Air Quality Network. 157

LHEM London Hybrid Exposure Model. 12, 13, 15, 121, 122, 126, 128, 129, 133–151, 172, 173, 177, 178, 181, 184–191, 194, 195, 209, 210, 215, 217

LTDS London Transport Demand Survey. 1, 11, 15, 91–93, 95, 96, 100, 103, 104, 117, 119, 172, 185, 189, 209, 210, 215, 216

LTDS-X London Transport Demand Survey - Expanded. 1, 88, 121, 122, 126–130, 147, 153, 171

LUR Land-use regression. 53, 54

Microsoft Access Microsoft Access. 93

MOE Margin of error. 193

NO_2 Nitrogen dioxide. 10–15, 21, 22, 26, 27, 29–31, 35, 39–41, 47, 48, 52, 54, 63, 65, 80–82, 88, 121, 123–127, 134–151, 153, 154, 179, 182, 183, 187, 192, 194, 197, 200, 202–204, 206, 213, 217

NO_x Oxides of nitrogen. 10, 19, 23, 31, 41, 47, 53, 54

O_3 Ozone. 10, 20, 31, 47, 81, 123, 183

OA Output Area. 217

ONS Office for National Statistics. 10, 67

OSM Open Street Map. 215

PAHs Polycyclic aromatic hydrocarbons. 20, 23

PDA Personal Digital Assistant. 58

PM Particulate matter (size agnostic). 15, 22, 23, 27, 28, 63, 68, 70, 77, 157

PM_{10} Particulate matter of a diameter of less than 10 micrometres. 10, 21–23, 27, 28, 47, 48, 52, 59, 69–71, 73, 75, 77, 83, 123, 142, 187

$PM_{2.5}$ Particulate matter of a diameter of less than 2.5 micrometres. 9, 10, 12, 13, 15, 21–23, 27, 28, 35–37, 39–42, 44–48, 50–52, 54, 56, 60, 62–64, 68–73, 75, 77, 88, 89, 121, 123, 126, 129, 134–143, 145–147, 149–160, 162–177, 179, 187, 189, 207, 211, 213, 214, 217

PNC Particle number count. 48

SLDF Spatial lines data frame. 197, 198

SO2$_2$ Sulphur dioxide. 20

SQL Structured Query Language. 11, 95, 103, 159, 217

TfL Transport for London. 1, 72, 87, 92, 100, 158, 159, 174, 210, 211, 216

UFPs Ultra-fine particles. 62, 74

UK United Kingdom. 23, 24, 33, 36, 37, 71, 77, 125, 209, 216

US United States. 47

US-EPA United States Environmental Protection Agency. 26, 27, 122

US-HEI United States Health Effects Institute. 28

USA United States of America. 40, 50, 63

VOCs Volatile organic compounds. 26, 54

WHO World Health Organisation. 15, 21, 22, 30, 34, 63

A. Code Listings

The following code contains pertinent examples of key R and SQL code used in this research

A.1 An example of in-vehicle modelling using a mass-balance approach

```
_## Code by James Smith, 22 September 2015
## Example below is for when a subject is on the DLR (vehicle mode 18)
## The output is the concentration at 7:57, two minutes after
## the journey started at 7:55
## The process is similar for other journey types
## Using NOX pollutant properties as proxy for journey_start_no2
## Replicating journey (ssid) 505004171020106

## pollutant properties
nox_d_vel           <- 0.432 ## deposition velocity
nox_res_r_zr        <- 0 ## nox resuspension rate

## vehicle properties
dep_sur             <- 74 ## internal surface area of the vehicle
no_ps               <- 70 ## number of passengers in vehicle
v_vol               <- 260 ## volume of the vehicle
exc_nat_zr          <- 5.2 ## natural hourly air exchange rate at start of the journey
fil_ef              <- 1 ## vehicle filter effeciency
exc_mec             <- 0 ## mechanical hourly air exchange rate
exc_nat             <- 10 ## natural hourly air exchange rate
no_acps             <- 30 ## number of active passengers in vehicle
passenger_surface_area    <- 2.92 ## surface area of a passenger

## Conditions
journey_start_no2   <- 120.2343160622219292  # Outdoor concentration at the start of the journey
journey_start_nox   <- 296.4997652826462043  # Outdoor concentration at the start of the journey
epoch_current_time  <- 1252052460            # Time at which the journey in question started (#7:57)
epoch_journey_start_time  <- 1252052400           # Time at which we want to know indoor concentration (#7:55)
current_no2         <- 49.1132098585253775   # Outdoor concentration taken from CMAQ-Urban
current_nox         <- 92.4817114798921030   # Outdoor concentration taken from CMAQ-Urban

lambda.win  <- exc_nat
n           <- fil_ef
lambda.hvac <- exc_mec
V.g         <- nox_d_vel
A.star      <- dep_sur + passenger_surface_area*no_ps
V           <- v_vol
no2.C.out   <- current_no2
nox.C.out   <- current_nox
Q           <- no_acps * nox_res_r_zr
lambda.win.star     <- 0.5
lambda.win0 <- exc_nat_zr
no2.C.out0  <- journey_start_no2
nox.C.out0  <- journey_start_nox
t           <- (epoch_current_time - epoch_journey_start_time+60)/3600

a       <-  lambda.win + n*lambda.hvac + V.g*(A.star/V)
a0      <-  lambda.win0 + n*lambda.hvac + V.g*(A.star/V)
```

```
49
50 no2_b   <-  lambda.win*no2.C.out + Q/V
51 no2_b0  <-  lambda.win0*no2.C.out0 + Q/V
52
53 nox_b   <-  lambda.win*nox.C.out + Q/V
54 nox_b0  <-  lambda.win0*nox.C.out0 + Q/V
55
56 no2_C.in0  <-  (lambda.win.star/(lambda.win.star + V.g*(A.star/V)))*no2.C.out0
57 nox_C.in0  <-  (lambda.win.star/(lambda.win.star + V.g*(A.star/V)))*nox.C.out0
58
59 no2_C.in    <-  (no2_C.in0 - no2_b0/a0)*exp(-a*t) + no2_b/a
60 nox_C.in    <-  (nox_C.in0 - nox_b0/a0)*exp(-a*t) + nox_b/a
61
62 #print(paste("a is", a))
63 #print(paste("a0 is", a0))
64 #print(paste("b is", b))
65 #print(paste("b0 is", b0))
66 #print(paste("C.in0 is", C.in0))
67 #print(paste("C.in is", C.in))
68 #print(t)
69 print(paste("no2 is ", round(no2_C.in,2)))
70 print(paste("nox is ", round(nox_C.in,2)))
```

A.2 Requesting a route using a bus from the TfL API

```
1
2
3
4
5  url <- paste("https://api.tfl.gov.uk/Journey/JourneyResults/", start_lat, ",", start_lon, "/to/", end_lat, ",", end
       _lon,
6              "?journeyPreference=LeastTime&mode=bus&app_id=", tfl_app_id, "&app_key=", tfl_app_key,
7              "&date=", format(start_time, '%Y%m%d'),
8              "&time=", format(start_time, '%H%M'),
9              sep="")
10
11 received <- RCurl::getURL(url)
12
13 temp <- try(fromJSON(received, simplify = FALSE))
14
15 bus_journey_leg_times <- data.frame(id = '',
16                                     duration = '',
17                                     stringsAsFactors = FALSE)
18
19 if (class(temp) != 'try-error') {
20
21 json_data <- fromJSON(received, simplify = FALSE)
22
23 } else { print('routing failed')}
24
25 if (json_data$`$type` == 'Tfl.Api.Presentation.Entities.JourneyPlanner.ItineraryResult, Tfl.Api.Presentation.
       Entities') {
26
27   temp_results_frame <- data.frame(id = numeric(),
28                                    lat = numeric(),
29                                    lon = numeric(),
30                                    mode = character(),
31                                    line = character(),
32                                    stringsAsFactors = FALSE)
33
34   for (r in 1:length(json_data$journeys[[2]]$legs)) {
35
36     if ('lineString' %in% names(json_data$journeys[[2]]$legs[[r]]$path)) {
37
38       ## Need the leg durations
39       bus_journey_leg_times[r,]     <- c(paste(3,r,sep=""), as.numeric(json_data$journeys[[2]]$legs[[r]]$duration))
40       line                          <-  json_data$journeys[[2]]$legs[[r]]$routeOptions[[1]]$name
41       linestring                    <-  json_data$journeys[[2]]$legs[[r]]$path$lineString
42       linestring                    <- gsub(" ", "", linestring, fixed=TRUE)
43       linestring                    <- gsub("[", "", linestring, fixed = TRUE)
44       linestring                    <- gsub("]", "", linestring, fixed = TRUE)
45       linestring                    <- unlist(strsplit(linestring, split = ","))
46
47       per_leg_results <- data.frame(id = numeric(),
48                                     lat = numeric(),
49                                     lon = numeric(),
50                                     mode = character(),
51                                     line = character(),
52                                     stringsAsFactors = FALSE)
53
54       l <- 2
55       m <- 1
56       for (k in 1:(length(linestring)/2)){
57         per_leg_results[k,] <- c(paste(3,r,sep=""), as.numeric(linestring[m]), as.numeric(linestring[l]), 'bus',
               line)
58         l <- l+2
59         m <- m+2
60       }
61
62       temp_results_frame <- rbindlist(list(temp_results_frame, per_leg_results), use.names=TRUE)
63       rm(per_leg_results)
64       temp_results_frame$id <-    as.numeric(temp_results_frame$id)
65       temp_results_frame$lat       <-    as.numeric(temp_results_frame$lat)
66       temp_results_frame$lon       <-    as.numeric(temp_results_frame$lon)
67       temp_results_frame           <-    data.frame(temp_results_frame)
68     } else {}
69
70   }
```

```
  temp_results_frame[temp_results_frame$line == "","line"] <- NA
  temp_results_frame[is.na(temp_results_frame$line),"mode"] <- 'walk'

  bus_result <- temp_results_frame
  rm(temp_results_frame, k, l, m, r, line, linestring, url, json_data,tfl_app_id, tfl_app_key,received,temp)
  print('Bus routing was completed using TFL Journey Planned API')
} else {
  print('routing failed')
}
```

A.3 Creating the geographical missclassification graphs and maps

```
library("RPostgreSQL")
library("ggplot2")
library("scales")
library("reshape")
library("scales")
library("ggmap")
library("sp")
library("maptools")
library("RColorBrewer")
library("rgdal")
library("gridExtra")

drv = dbDriver("PostgreSQL")
con = dbConnect(drv, dbname="james_traffic", user="james", password="XXXXX", host="localhost")

locations <- dbGetQuery(con, paste("
SELECT    person.ppid,
          st_x(st_setsrid(st_makepoint(hhose::numeric, hhosn::numeric),27700)) AS x,
          st_y(st_setsrid(st_makepoint(hhose::numeric, hhosn::numeric),27700)) AS y
FROM      person
LEFT JOIN household
ON        person.phid = household.hhid
WHERE   ppiwt::numeric > 0 AND bad_flag IS NULL
"))

## Set a working directory to output the graphs too
setwd("/home/james/mounts/James/PhD/9 - Dynamic Comparison Chapter/Outputs/geographical_missclassification/")

## import the missclassification data I've already ran
load("~/mounts/James/PhD/9 - Dynamic Comparison Chapter/Outputs/address_point_v_lhem/results.Rda")

## now need to link the two on PPID
new_results <- merge(locations, results, by='ppid')

rm(results)
rm(locations)

new_results$missclassification_percentage_pm25 <- 100 * ((new_results$lhem_pm25 - new_results$household_pm25) / new
    _results$household_pm25)
new_results$missclassification_percentage_no2 <- 100 * ((new_results$lhem_no2 - new_results$household_no2) / new_
    results$household_no2)

london <- readOGR(dsn = ".", layer = "london")
london <- fortify(london, region="name")

## First make a cumulative distribution plot of the missclassification percentages

plot1 <- ggplot(new_results, aes(missclassification_percentage_pm25)) + stat_ecdf(size=2, colour="red") +
  theme(axis.line = element_line(colour = "black")) +
  theme(axis.line = element_line(colour="black"),
        axis.text=element_text(size=20, color="black"),
        axis.title=element_text(size=20, color="black"),
        plot.title=element_text(size=20, colour="black"),
        legend.text=element_text(size=20, colour = "black"),
        legend.title=element_blank(),
        legend.justification=c(1,1),
        legend.position=c(1,1)) +
  labs(#title=expression(paste("Mean daily exposure to PM"[2.5], " (", mu, "g m" ^ "-3", "), (LHEM v. Address-point
      )")),
    x=expression(paste("PM"[2.5], " Missclassification %")),
    y="Cumulative percentage of subjects")

plot2 <- ggplot(new_results, aes(missclassification_percentage_pm25)) + stat_ecdf(size=2, colour="blue") +
  theme(axis.line = element_line(colour = "black")) +
  theme(axis.line = element_line(colour="black"),
        axis.text=element_text(size=20, color="black"),
        axis.title=element_text(size=20, color="black"),
        plot.title=element_text(size=20, colour="black"),
        legend.text=element_text(size=20, colour = "black"),
        legend.title=element_blank(),
```

```
        legend.justification=c(1,1),
        legend.position=c(1,1)) +
  labs(#title=expression(paste("Mean daily exposure to PM"[2.5], " (", mu, "g m" ^ "-3", "), (LHEM v. Address-point)")),
    x=expression(paste("NO"[2], " missclassification %")),
    y="")

pdf("cumulative_missclass_dist.pdf", width=14, height=6.4)
pushViewport(viewport(layout = grid.layout(1, 2)))
print(plot1, vp = viewport(layout.pos.row = 1, layout.pos.col = 1))
print(plot2, vp = viewport(layout.pos.row = 1, layout.pos.col = 2))
dev.off()

## Now a map of the people with an increase

plot1 <- ggplot(data = new_results[new_results$missclassification_percentage_no2 > 0, ], aes(x = x, y = y)) +
      geom_polygon(data = london, aes(x = long, y = lat, group = group), fill="grey", color = "black") +
  geom_point(size=2.5, colour="red") +
      theme(panel.grid = element_blank(),
        axis.text = element_blank(),
        axis.title =  element_text(size=20, color="black"),
        legend.title = element_blank(),
        panel.background = element_blank(),
        axis.ticks = element_blank()) +
  labs(title="", x=expression(paste("PM"[2.5])), y="")

plot2 <- ggplot(data = new_results[new_results$missclassification_percentage_pm25 > 0, ], aes(x = x, y = y)) +
  geom_polygon(data = london, aes(x = long, y = lat, group = group), fill="grey", color = "black") +
  geom_point(size=2.5, colour="blue") +
  theme(panel.grid = element_blank(),
        axis.text = element_blank(),
        axis.title =  element_text(size=20, color="black"),
        legend.title = element_blank(),
        panel.background = element_blank(),
        axis.ticks = element_blank()) +
  labs(title="", x=expression(paste("NO"[2])), y="")

pdf("address_lhem_increases.pdf", width=14, height=6.4)
pushViewport(viewport(layout = grid.layout(1, 2)))
print(plot1, vp = viewport(layout.pos.row = 1, layout.pos.col = 1))
print(plot2, vp = viewport(layout.pos.row = 1, layout.pos.col = 2))
dev.off()
```

A.4 Creating a London Underground GIS file

```
 DROP VIEW IF EXISTS tube_ext;

--CREATE underground_stations table

DROP TABLE IF EXISTS underground_stations;

CREATE TABLE underground_stations(
name VARCHAR,
description VARCHAR,
x INTEGER,
y INTEGER);

-- Get the data from the CSV file
COPY underground_stations from '/home/james/mounts/James/PhD/2 - Routing/Tube/Raw Data/Underground_Stations_with_
    missing.csv' DELIMITERS ',' HEADER CSV;

-- Add an ID column and make it the primary key
ALTER TABLE underground_stations
ADD COLUMN id SERIAL PRIMARY KEY;

-- Add a Geom column

SELECT AddGeometryColumn ('underground_stations','the_geom',27700,'POINT',2);

-- Populate the Geom column from the easting and northing attributes

UPDATE underground_stations
SET the_geom = st_setsrid(st_makepoint("x", "y"),27700)
WHERE x is not null;

-- Now create an underground_routes table

DROP TABLE IF EXISTS underground_routes;

CREATE TABLE underground_routes(
shortname VARCHAR,
longname VARCHAR,
line VARCHAR,
in_order INTEGER,
section INTEGER,
ground_level numeric,
northbound numeric,
southbound numeric,
eastbound numeric,
westbound numeric,
inputted text
);

-- Get the data from the CSV file
-- Y:\James\PhD\2 - Routing\Tube\Raw Data
COPY underground_routes from '/home/james/mounts/James/PhD/2 - Routing/Tube/Raw Data/Lines_and_segments_created_
    with_depths.csv' DELIMITERS ',' CSV HEADER;

-- Add an ID column and make it the primary key
ALTER TABLE underground_routes
ADD COLUMN id SERIAL PRIMARY KEY;

-- Add a Geom column

SELECT AddGeometryColumn ('underground_routes','the_geom',27700,'POINT',2);

-- Join the geom of the points and the order and routes together into the underground_routes table.

UPDATE underground_routes
SET the_geom = underground_stations.the_geom
FROM underground_stations
WHERE underground_routes.longname = underground_stations.name;

-- Create a table with the manual coordinates of the tube stations that don't come automatically from TFL

DROP TABLE IF EXISTS manual_stations;

CREATE TABLE manual_stations(
```

```
id SERIAL PRIMARY KEY,
shortname VARCHAR,
longname VARCHAR
);

-- ADD the geometry columns to this table

SELECT AddGeometryColumn ('manual_stations','the_geom',27700,'POINT',2);

-- ADD THE TABLE MANUALLY TO THIS TABLE

INSERT INTO manual_stations (shortname, longname, the_geom)
VALUES
('Wood Lane', 'Wood Lane Station',  ST_Transform(st_geomfromtext('POINT(-0.2242 51.5098)', 4326), 27700)),
('Heathrow Terminal 5', 'Heathrow Terminal 5 Station',  ST_Transform(st_geomfromtext('POINT(-0.488 51.4723)', 4326), 27700)),
('Langdon Park', 'Langdon Park Station', ST_Transform(st_geomfromtext('POINT(-0.014 51.515)', 4326), 27700)),
('Star Lane', 'Star Lane Station', ST_Transform(st_geomfromtext('POINT(0.0042 51.5207)', 4326), 27700)),
('Abbey Road', 'Abbey Road Station', ST_Transform(st_geomfromtext('POINT(0.004 51.532)', 4326), 27700)),
('Stratford High Street', 'Stratford High Street Station', ST_Transform(st_geomfromtext('POINT(-0.0006 51.5379 )', 4326), 27700)),
('Stratford International', 'Stratford International Station', ST_Transform(st_geomfromtext('POINT(-0.0086 51.5448)', 4326), 27700)),
('West Silvertown', 'West Silvertown Station', ST_Transform(st_geomfromtext('POINT(0.0225 51.502778)', 4326), 27700)),
('Pontoon Dock', 'Pontoon Dock Station', ST_Transform(st_geomfromtext('POINT(0.031944 51.502222)', 4326), 27700)),
('London City Airport', 'London City Airport Station', ST_Transform(st_geomfromtext('POINT(0.048889 51.503611)', 4326), 27700)),
('King George V', 'King George V Station', ST_Transform(st_geomfromtext('POINT(0.062778 51.501972)', 4326), 27700)),
('Woolwich Arsenal', 'Woolwich Arsenal Station', ST_Transform(st_geomfromtext('POINT(0.069 51.49)', 4326), 27700))
;

-- Take the missing geoms from the manual_stations table and move them into the main underground_routes table

UPDATE underground_routes
SET the_geom = manual_stations.the_geom
FROM manual_stations
WHERE underground_routes.longname = manual_stations.longname;

-- STAGE TWO: MAKE THE LINES/ROUTES
-- Need start and end points for each of the lines.

DROP TABLE IF EXISTS temp;

CREATE TABLE  temp AS (
SELECT    underground_routes.line,
    underground_routes.section,
    underground_routes.shortname AS start_station,
    underground_routes.in_order AS start_node,
    underground_routes.the_geom AS start_node_geom,
        a.shortname AS end_station,
    a.in_order AS end_node,
    a.the_geom AS end_node_geom
FROM    underground_routes a
INNER JOIN  underground_routes ON a.section = underground_routes.section
WHERE     underground_routes.line = a.line
AND     underground_routes.in_order <> a.in_order
ORDER BY  underground_routes.line, section, start_node, end_node
);

--SELECT THE ABOVE TABLE AND MAKE A LINE BETWEEN EACH SET OF POINTS

DROP TABLE IF EXISTS underground_routes_processed CASCADE;

CREATE TABLE underground_routes_processed AS (
SELECT  *,
  st_makeline(start_node_geom, end_node_geom) AS edge
FROM  temp
WHERE   temp.end_node - temp.start_node = '1'
OR  temp.end_node - temp.start_node = '-1'
);

-- Now start to make the routing network

CREATE VIEW tube_ext AS
SELECT *, st_startpoint(edge), st_endpoint(edge)
FROM underground_routes_processed;
```

```
144
145 -- Create a table of nodes AKA change points/junctions. Remove duplicates.
146
147 DROP TABLE IF EXISTS tube_node;
148
149 CREATE TABLE tube_node AS
150    SELECT row_number() OVER (ORDER BY foo.p)::integer AS id,
151           foo.p AS the_geom
152    FROM (
153       SELECT DISTINCT tube_ext.st_startpoint AS p FROM tube_ext
154       UNION
155       SELECT DISTINCT tube_ext.st_endpoint AS p FROM tube_ext
156    ) foo
157    GROUP BY foo.p;
158
159 -- Link the nodes to the lines
160
161 DROP TABLE IF EXISTS tube_network;
162
163 CREATE TABLE tube_network AS
164    SELECT a.*, b.id as start_id, c.id as end_id
165    FROM tube_ext AS a
166       JOIN tube_node AS b ON a.st_startpoint = b.the_geom
167       JOIN tube_node AS c ON a.st_endpoint = c.the_geom;
168
169 -- Add a serial and make it the primary key
170
171 ALTER TABLE tube_network ADD COLUMN id SERIAL;
172 ALTER TABLE tube_network ADD CONSTRAINT tube_network_pk PRIMARY KEY (id);
```

A.5 Creating the central line map

```
rm(list=ls())

## Load libraries
library("RPostgreSQL")
library("ggplot2")
library("scales")
library("reshape")
library("grid")
library("devtools")
library("openair")
library("gdata")

setwd("/home/james/mounts/James/PhD/10 - Tube Monitoring Chapter/Results")

## Connect to PostgreSQL database where data is stored
drv = dbDriver("PostgreSQL")
con = dbConnect(drv, dbname="tube_monitoring", user="james", password="", host="10.0.4.240")

## Get daily background averages for London from OpenAir. for the days we did monitoring on.
background_pm25 <- importKCL(site = "kc1", year = c(2014,2015,2016), pollutant = "pm25", met = FALSE,
                             units = "mass", extra = FALSE)
background_pm25 <- data.frame(background_pm25, day = as.Date(format(background_pm25$date)))
background_pm25 <- aggregate(pm25 ~ day, background_pm25, mean)

## Extract the data from database. Note the CASE statement below. it's because want the raw PM25 data to do our own
      scaling, but the scaled version of other pollutants.
tube_data <- dbGetQuery(con, "
                        WITH  station_depths AS (
                        SELECT    station_depths_import.station_name,
                        station_depths_import.line,
                        station_depths_import.platform_depth,
                        station_info.shortname,
                        station_info.the_geom
                        FROM    station_depths_import
                        LEFT JOIN ( SELECT    shortname,
                        line,
                        the_geom
                        FROM    station_geom_depth
                        GROUP BY  shortname,
                        line,
                        the_geom) AS station_info
                        ON    station_depths_import.station_name  = station_info.shortname
                        AND   station_depths_import.line    = station_info.line
                        )
                        SELECT    tube_pollution_mapping.species,
                        tube_pollution_mapping.environment,
                        tube_pollution_mapping.date_time,
                        CASE WHEN tube_pollution_mapping.species = 'PM25' THEN tube_pollution_mapping.concentration
                        ELSE tube_pollution_mapping.scaled_concentration
                        END AS concentration,
                        tube_pollution_mapping.tube_diary_stop,
                        tube_pollution_mapping.line,
                        station_depths.platform_depth
                        FROM    tube_pollution_mapping
                        LEFT JOIN station_depths
                        ON    tube_pollution_mapping.tube_diary_stop  =   station_depths.station_name
                        AND   tube_pollution_mapping.line   = station_depths.line
                        ORDER BY  tube_pollution_mapping.date_time,
                        tube_pollution_mapping.environment,
                        tube_pollution_mapping.species
                        ")

## Link the tube data to the background air quality data
tube_data <- data.frame(tube_data, day = as.Date(format(tube_data$date_time)))
tube_data <- merge(tube_data, background_pm25, by="day", all.x=TRUE)

# Find that concentration for 1 Feb 2016 is missing from the London Air background PM2.5 data. Need some numbers
      for that. So going to use the data from 8 Feb 2016 instead. It's the same day of the week, the week after.
tube_data[is.na(tube_data$pm25),]$pm25 <- background_pm25[background_pm25$day == '2016-02-08',]$pm25

# Now do the correction process. Create new field to put the data in.
tube_data$corrected_concentration <- as.numeric('')

```

```
## Now adjust the data. Using the background concentration data, adjust the tube concentration data using the
      regression formula determined by Dave Green and Ben Barratt
for (i in 1:nrow(tube_data)) {
  if (tube_data[i,]$concentration > tube_data[i,]$pm25/0.44 & tube_data[i,]$species == 'PM25') {
    tube_data[i,]$corrected_concentration <- (tube_data[i,]$pm25 + (tube_data[i,]$concentration - tube_data[i,]$
         pm25/0.44) * 1.82)
  } else
  {
    if (tube_data[i,]$concentration <= tube_data[i,]$pm25/0.44 & tube_data[i,]$species == 'PM25') {
      tube_data[i,]$corrected_concentration <-tube_data[i,]$concentration * 0.44
    } else
    {}
  }
}


tube_data$corrected_concentration[is.na(tube_data$corrected_concentration)] <- tube_data$concentration[is.na(tube_
      data$corrected_concentration)]

## Make a data frame of tube lines and colours for plotting
colours_lines <- data.frame(line = c("Victoria","Piccadilly","Northern","Circle","Jubilee","District","Bakerloo","
      Metropolitan","Docklands Light Railway","Central","Hammersmith & City"),
                            colour = c("#0099CC","#000099","#000000","#FFCC00","#868F98","#006633","#996633","
                                 #660066","#009999","#CC3333", "#CC9999"),
                            stringsAsFactors = FALSE)

## Clean up some data don't need anymore
rm(background_pm25, con, drv, i)

## Aggregate the data that we're going to need so it's suitable for plotting
map_plot <- aggregate(corrected_concentration ~ tube_diary_stop + line + platform_depth, tube_data[tube_data$
      species == 'PM25' & tube_data$line == 'Central' & tube_data$environment == 'CAR',], mean)

## Rename the variables
names(map_plot)[names(map_plot) == 'tube_diary_stop'] <- 'station'
names(map_plot)[names(map_plot) == 'platform_depth'] <- 'depth'
names(map_plot)[names(map_plot) == 'corrected_concentration'] <- 'pm25'

## We don't want to use the lat and long to plot, as we want to create a Beck style map of the tube. So we are
      going to give the stations pseudo coordinates to plot it how we like.
map_plot$fake_x <- NA
map_plot$fake_y <- NA

## Having planned this out on graph paper, the coordinates for the stations are now entered
map_plot[map_plot$line == 'Central' & map_plot$station == 'West Ruislip',c("fake_x", "fake_y")]            <- c
      (1,2)
map_plot[map_plot$line == 'Central' & map_plot$station == 'Ruislip Gardens',c("fake_x", "fake_y")]         <- c
      (2,2)
map_plot[map_plot$line == 'Central' & map_plot$station == 'South Ruislip',c("fake_x", "fake_y")]           <- c
      (3,2)
map_plot[map_plot$line == 'Central' & map_plot$station == 'Northolt',c("fake_x", "fake_y")]                <- c
      (4,2)
map_plot[map_plot$line == 'Central' & map_plot$station == 'Greenford',c("fake_x", "fake_y")]               <- c
      (5,2)
map_plot[map_plot$line == 'Central' & map_plot$station == 'Perivale',c("fake_x", "fake_y")]                <- c
      (6,2)
map_plot[map_plot$line == 'Central' & map_plot$station == 'Hanger Lane',c("fake_x", "fake_y")]             <- c
      (7,2)
map_plot[map_plot$line == 'Central' & map_plot$station == 'North Acton',"fake_x"]                          <- 8
map_plot[map_plot$line == 'Central' & map_plot$station == 'North Acton',"fake_y"]                          <- 2
map_plot[map_plot$line == 'Central' & map_plot$station == 'East Acton',c("fake_x", "fake_y")]              <- c
      (9,2)
map_plot[map_plot$line == 'Central' & map_plot$station == 'White City',c("fake_x", "fake_y")]              <- c
      (10,2)
map_plot[map_plot$line == 'Central' & map_plot$station == "Shepherd's Bush",c("fake_x", "fake_y")]         <- c
      (11,2)
map_plot[map_plot$line == 'Central' & map_plot$station == 'Holland Park',c("fake_x", "fake_y")]            <- c
      (12,2)
map_plot[map_plot$line == 'Central' & map_plot$station == 'Notting Hill Gate',c("fake_x", "fake_y")]       <- c
      (13,2)
map_plot[map_plot$line == 'Central' & map_plot$station == 'Queensway',c("fake_x", "fake_y")]               <- c
      (14,2)
map_plot[map_plot$line == 'Central' & map_plot$station == 'Lancaster Gate',c("fake_x", "fake_y")]          <- c
      (15,2)
map_plot[map_plot$line == 'Central' & map_plot$station == 'Marble Arch',c("fake_x", "fake_y")]             <- c
      (16,2)
map_plot[map_plot$line == 'Central' & map_plot$station == 'Bond Street',c("fake_x", "fake_y")]             <- c
      (17,2)
```

```
127 map_plot[map_plot$line == 'Central' & map_plot$station == 'Oxford Circus',c("fake_x", "fake_y")]           <- c
      (18,2)
128 map_plot[map_plot$line == 'Central' & map_plot$station == 'Tottenham Court Road',c("fake_x", "fake_y")]    <- c
      (19,2)
129 map_plot[map_plot$line == 'Central' & map_plot$station == 'Holborn',c("fake_x", "fake_y")]                 <- c
      (20,2)
130 map_plot[map_plot$line == 'Central' & map_plot$station == 'Chancery Lane',c("fake_x", "fake_y")]           <- c
      (21,2)
131 map_plot[map_plot$line == 'Central' & map_plot$station == "St. Paul's", c("fake_x", "fake_y")]            <- c
      (22,2)
132 map_plot[map_plot$line == 'Central' & map_plot$station == 'Bank',c("fake_x", "fake_y")]                    <- c
      (23,2)
133 map_plot[map_plot$line == 'Central' & map_plot$station == 'Liverpool Street',c("fake_x", "fake_y")]        <- c
      (24,2)
134 map_plot[map_plot$line == 'Central' & map_plot$station == 'Bethnal Green',c("fake_x", "fake_y")]           <- c
      (25,2)
135 map_plot[map_plot$line == 'Central' & map_plot$station == 'Mile End',c("fake_x", "fake_y")]               <- c
      (26,2)
136 map_plot[map_plot$line == 'Central' & map_plot$station == 'Stratford',c("fake_x", "fake_y")]              <- c
      (27,2)
137 map_plot[map_plot$line == 'Central' & map_plot$station == 'Leyton',c("fake_x", "fake_y")]                 <- c
      (28,2)
138 map_plot[map_plot$line == 'Central' & map_plot$station == 'Leytonstone',"fake_x"]                         <- 29
139 map_plot[map_plot$line == 'Central' & map_plot$station == 'Leytonstone',"fake_y"]                         <- 2
140 map_plot[map_plot$line == 'Central' & map_plot$station == 'Wanstead',c("fake_x", "fake_y")]               <- c
      (30,2)
141 map_plot[map_plot$line == 'Central' & map_plot$station == 'Redbridge',c("fake_x", "fake_y")]             <- c
      (31,2)
142 map_plot[map_plot$line == 'Central' & map_plot$station == 'Gants Hill',c("fake_x", "fake_y")]             <- c
      (32,2)
143 map_plot[map_plot$line == 'Central' & map_plot$station == 'Newbury Park',c("fake_x", "fake_y")]           <- c
      (33,2)
144 map_plot[map_plot$line == 'Central' & map_plot$station == 'Barkingside',c("fake_x", "fake_y")]           <- c
      (34,2)
145 map_plot[map_plot$line == 'Central' & map_plot$station == 'Fairlop',c("fake_x", "fake_y")]               <- c
      (35,2)
146 map_plot[map_plot$line == 'Central' & map_plot$station == 'Hainault',c("fake_x", "fake_y")]              <- c
      (36,2)
147 #map_plot[map_plot$line == 'Central' & map_plot$station == 'West Acton',c("fake_x", "fake_y")]             <- c
      (6,1)
148 #map_plot[map_plot$line == 'Central' & map_plot$station == 'Ealing Broadway',c("fake_x", "fake_y")]        <- c
      (5,1)
149
150 ## Now categorise the continuous depth data for plotting as different colours.
151
152 map_plot$depth_categorised <- NA
153 map_plot[map_plot$depth < 0,]$depth_categorised  <- 'above ground (>0m)'
154 map_plot[map_plot$depth < 10 & map_plot$depth > 0,]$depth_categorised  <- 'shallow (0-10m)'
155 map_plot[map_plot$depth < 20 & map_plot$depth > 10,]$depth_categorised  <- 'medium (10-20m)'
156 map_plot[map_plot$depth > 20,]$depth_categorised  <- 'deep (>20m)'
157
158 ## Make into factors for plotting
159 map_plot$depth_categorised <- as.factor(map_plot$depth_categorised)
160 map_plot$depth_categorised <- factor(map_plot$depth_categorised, levels = c("above ground (>0m)", "shallow (0-10m)"
      ,
161                                                                              "medium (10-20m)", "deep (>20m)"))
162
163 ## Rescale the PM2.5 to be between 0 and 1 so that we can visualise it better
164 map_plot[map_plot$line == 'Central',"scaled_pm25"] <- rescale(map_plot[map_plot$line == 'Central',]$pm25, to=c(0,1)
      )
165
166 ## Make the plot and save as a PNG
167 png("fake_central_line_pm25_no_spur.png", width =1000, height = 400, units="px")
168 ggplot(map_plot[map_plot$line == 'Central',], aes(fake_x, fake_y, label = station, colour = depth_categorised)) +
169       geom_segment(data = map_plot[map_plot$line == 'Central',], aes(x = fake_x, y = fake_y,
170                                                                      xend = fake_x, yend = fake_y +scaled_pm25),
171                    colour = "black", size = 3, lineend="round", alpha=0.4) +
172       geom_line(data = data.frame(fake_x = c(1,36), fake_y = 2, station = NA, depth_categorised = NA),
173                 aes(x = fake_x1=,enorsuv, y = fake_y), size = 2, colour = "red") +
174      # geom_line(data = data.frame(fake_x = c(5,8), fake_y = 1, station = NA, depth_categorised = NA),
175      #           aes(x = fake_x, y = fake_y), size = 2, colour = "red") +
176      # geom_line(data = data.frame(fake_x = 8, fake_y = c(1,2), station = NA, depth_categorised = NA),
177      #            aes(x = fake_x, y = fake_y), size = 2, colour = "red") +
178            geom_point(size = 5.5, colour = "black") +
179       geom_point(size = 5) +
180            scale_colour_manual(values = c("white", "pink", "red", "black"), guide = guide_legend(title = "Station
                  depth")) +
181       geom_text(angle = 55, hjust = 1, nudge_y = -0.05, size = 5, colour = "black") +
```

```
        annotate(geom = "text", x = 0, y = 2.5, label = "PM[2.5] ~mu~g/m^3", color = "black", vjust = 0.1,
                 angle = 90, parse = TRUE, size = 6) +
       annotate("segment", x = 0.5, xend = 0.5, y = 2, yend = 3,
                 colour = "black") +
       annotate(geom = "text", x = 0.25, y = 2, label = "0", color = "black", hjust = 0.8,
                 angle = 0, parse = TRUE, size = 6) +
       annotate(geom = "text", x = 0, y = 3, label = "500", hjust = 0.8, color = "black",
                 angle = 0, parse = TRUE, size = 6) +
       ylim(1,3) +
       theme(axis.title = element_blank(),
             axis.text = element_blank(),
             axis.ticks=element_blank(),
             legend.position = c(0.86,0.86),
             legend.text = element_text(size = 12),
             legend.title = element_text(size = 12),
             legend.key.size = unit(1, "cm"),
             legend.background = element_rect(color = "black"),
             panel.grid = element_blank(),
             panel.background = element_rect(fill = "white"),
             plot.margin = unit(c(1,1,1,1), "cm"))
dev.off()
```

A.6 Creating example distribution plots for the LHEM

```
library(ggplot2)

minutes           <- seq(1:100)
low               <- c(runif(33, 3, 7))
medium            <- c(runif(34, 40, 60))
high              <- c(runif(33, 3, 7))
concentrations    <- c(low, medium, high)

data        <- data.frame(minutes, concentrations)

plot <- ggplot(data, aes(minutes, concentrations)) +
      geom_line() +
      theme(axis.text = element_text(size = 14, colour = "black"),
            axis.title = element_text(size = 14, colour = "black")
            ) +
      annotate("rect", xmin = 0, xmax = 100, ymin = 0, ymax = 6,
               alpha = .2, colour="blue") +
      annotate("rect", xmin = 0, xmax = 100, ymin = 8, ymax = 60,
                alpha = .2, colour="green") +
      geom_hline(aes(yintercept=7), colour="red") +
      xlab("Mintes of journey on tube") +
      ylab("PM2.5 concentation") +
      annotate("text", 50, 20, label = "x 1.82", size = 6) +
      annotate("text", 50, 3, label = "x 0.44", size = 6)

setwd("/home/james/mounts/James/PhD/10 - Tube Monitoring Chapter/Results")

ggsave(plot, file="tube_correction_example.png", width=10, height=10, units = "cm")
```

A.7 Creating August-September 9am air quality data

```
rm(list = ls())

library(lubridate)
library(raster)
library(M3)
library(readxl)
library(rgdal)
library(RCurl)
library(lattice)
library(rasterVis)
library(ggplot2)
library(jsonlite)

latlong    <- "+init=epsg:4326"
ukgrid     <- "+init=epsg:27700"
google     <- "+init=epsg:3857"
cmaqurban  <- "+proj=lcc +lat_1=35 +lat_2=65 +lat_0=52 +lon_0=10 +a=6370000 +b=6370000"
cmaq       <- "+proj=lcc +lat_1=46 +lat_2=46 +lat_0=46 +lon_0=17 +a=6370000 +b=6370000"

setwd("/home/james/Desktop")

## want to make annual average weekday, 9am to 10am for August and September

## Get list of the data
days_needed         <- data.frame(days = seq.POSIXt(as.POSIXct('2016-08-01'), as.POSIXct('2016-09-30'), 'days'))

days_needed$weekday <- weekdays(days_needed$days)

days_needed         <- days_needed[!days_needed$weekday %in% c('Saturday', 'Sunday'),]

days_needed$file    <-  paste0("001.cmaqurban.",
                               year(days_needed$days), format(days_needed$days, '%m'), format(days_needed$days, '%d'),
                               ".1.ERG20m.NO2.CONC.ncf")

for (i in 1:nrow(days_needed)) {

  print(paste0("Getting day ",  days_needed$days[i]))

  system(paste0("sshpass -p '***' scp james@10.0.4.226:/mnt/C3filestore/cope/COPE_VBS_2016_scem/cmaqurban/*/",
                days_needed$file[i],
                " ."))

  print("Got it")

  ncf_no2_file <-  days_needed$file[i]

  cmaq_no2_data <- brick(ncf_no2_file, var = "NO2", values=T)

  print('Sorting out the parameters')

  xmin(cmaq_no2_data)          <- get.grid.info.M3(ncf_no2_file)$x.orig
  ymin(cmaq_no2_data)          <- get.grid.info.M3(ncf_no2_file)$y.orig
  ymax(cmaq_no2_data)          <- get.grid.info.M3(ncf_no2_file)$y.orig +
    get.grid.info.M3(ncf_no2_file)$y.cell.width *
    get.grid.info.M3(ncf_no2_file)$nrows
  xmax(cmaq_no2_data)          <- get.grid.info.M3(ncf_no2_file)$x.orig +
    get.grid.info.M3(ncf_no2_file)$x.cell.width *
    get.grid.info.M3(ncf_no2_file)$ncols

  proj4string(cmaq_no2_data)  <- CRS(get.proj.info.M3(ncf_no2_file))

  res(cmaq_no2_data)          <- 20 # (It's a 20m by 20m grid file)

  print('Getting the 9am to 10am data')

  cmaq_no2_data               <- cmaq_no2_data$X10

  print('Adding it to the previous one')

  if (i == 1) {
    aug_sept_layer            <- cmaq_no2_data
  } else {
```

```
    aug_sept_layer            <- aug_sept_layer + cmaq_no2_data

    print('Thats this one done. Going to delete it now.')

    system(paste0("rm ", days_needed$file[i]))
  }

  rm(cmaq_no2_data)

}

system(paste0("rm ", days_needed$file[1]))

aug_sept_layer       <- aug_sept_layer / nrow(days_needed)

rm(ncf_no2_file, i)

writeRaster(aug_sept_layer, overwrite=T, '/home/james/mounts/James/PhD/11 - Evaluation Chapter/pollutant_files/no2_
    london_aug_sep.tif', format = 'GTiff')
writeRaster(my_area_no2, '/home/james/mounts/James/PhD/11 - Evaluation Chapter/pollutant_files/no2_cycling_area_aug
    _sep.tif', format = 'GTiff')
```